DOCUMENT RESUME

ED 196 946 TM 810 069

AUTHOR Stalford, Charles B., Ed.
TITLE Testing and Evaluation in Schools: Practitioners' Views.
INSTITUTION National Inst. of Education (ED), Washington, D.C.
PUB DATE Oct 80
NOTE 124p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Achievement Tests; *Administrator Attitudes; Decision Making; *Educational Assessment; *Educational Change; Elementary Secondary Education; *Program Evaluation; School Districts; Special Education; Standardized Tests; State School District Relationship; *Teacher Attitudes; *Testing

ABSTRACT

This volume is a compendium of local level practitioners' papers about testing and evaluation issues in the schools. Excerpts from 13 papers and a summary of the proceedings of a conference sponsored by the National Institute of Education's Testing, Assessment and Evaluation Division are included. Titles of the papers are: (1) Testing Concerns of a Special Educator; (2) Florida's Standardized Testing Program: A Tool or Weapon?; (3) Making Reading Achievement Tests Work for the Inner-City Student; (4) Teachers on Testing; (5) Standardized Testing in Elementary School: A Practitioner's Perspective on Several Significant Testing and Evaluation Issues; (6) An Elementary School Principal's View of Standardized Testing; (7) Coordinating Testing, Evaluation, and Decisionmaking at the Local Level; (8) Evaluation of a Planned Change Effort; (9) Responding to Conflicting Evaluation Demands; (10) Problems of Measuring Achievement and How They Are Being Addressed in the Portland, Oregon, Public Schools; (11) Some Problems of Evaluation in Large School Districts; (12) What It Takes To Win: Factors in the Utilization of Evaluation Findings for Educational Improvement; and (13) Producing Quality Program Evaluation in Education and Using It: The Washington, D.C., Experience. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

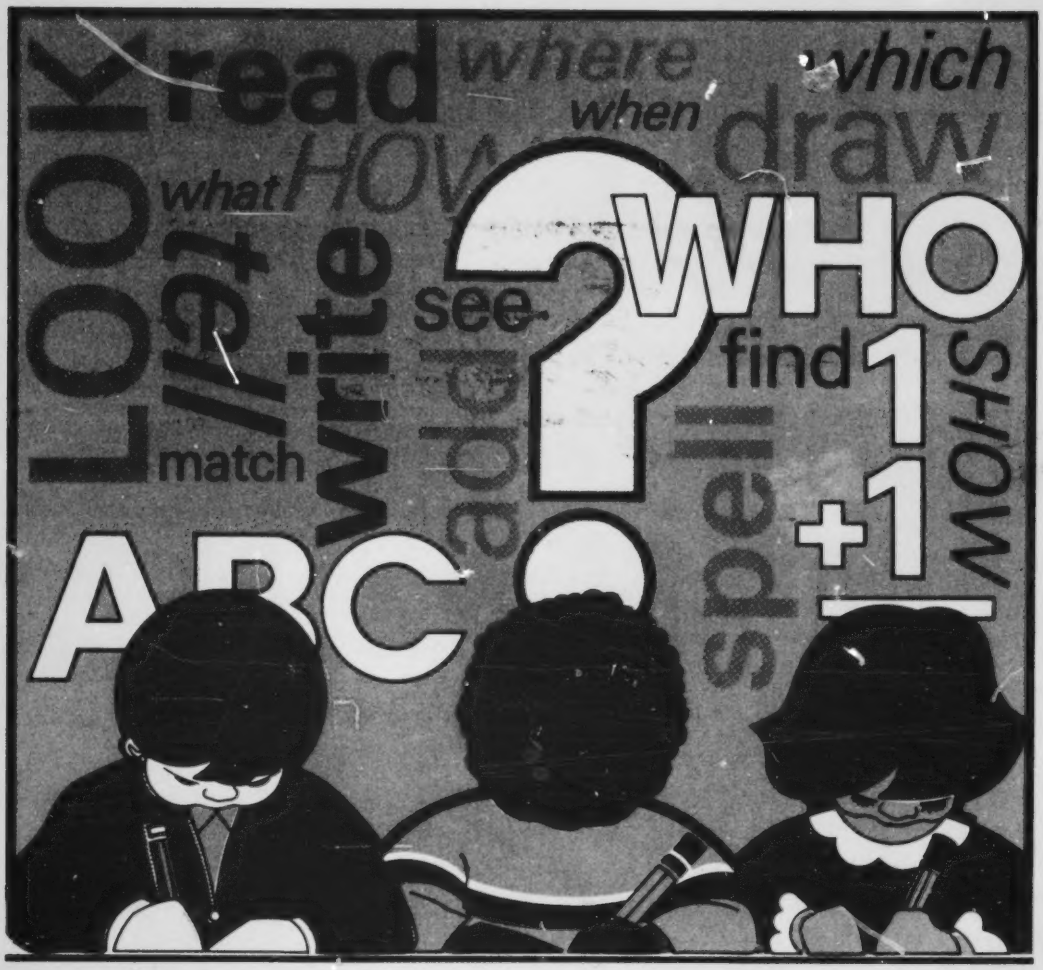
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED196946

TESTING AND EVALUATION IN SCHOOLS:

Practitioners' Views



TM 810 069

The National Institute of Education
U.S. Department of Education
Washington, D.C. 20208

Testing and Evaluation in Schools: Practitioners' Views

Charles B. Stalford, Editor

U.S. DEPARTMENT OF EDUCATION
Shirley M. Hufstedler, Secretary
Steven A. Minter, Under Secretary

OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT
F. James Rutherford, Assistant Secretary

NATIONAL INSTITUTE OF EDUCATION
Michael Timpane, Director
Lois-ellin Datta, Associate Director for Teaching and Learning
Jeffry Schiller, Assistant Director for Testing, Assessment, and Evaluation

October 1980

JAN 2 6 1981

Contents

Introduction 1

Teachers

1. Testing Concerns of a Special Educator

   Judith Singleton 9

2. Florida's Standardized Testing Program: A Tool or Weapon?

   William Moore 19

3. Making Reading Achievement Tests Work for the Inner-City Student

   Edward J. Cypress 27

4. Teachers on Testing

   Myrna Cooper and Maurice Leiter 35

Principals

5. Standardized Testing in Elementary School: A Practitioner's Perspective on Several Significant Testing and Evaluation Issues

   Parker Damon 45

6. An Elementary School Principal's View of Standardized Testing

   Luis Mercado 53

7. Coordinating Testing, Evaluation, and Decisionmaking at the Local Level

   Blas M. Garza, Jr. 59

8. Evaluation of a Planned Change Effort

   M. Claradine Johnson 67

Research and Evaluation Staff

9. Responding to Conflicting Evaluation Demands

   Michael H. Kean 77

10. Problems of Measuring Achievement and How They Are Being Addressed in the Portland, Oregon, Public Schools

Victor W. Doherty 87

11. Some Problems of Evaluation in Large School Districts

Ronald E. Banks 93

12. What It Takes To Win: Factors in the Utilization of Evaluation Findings for Educational Improvement

Freda M. Holley 101

13. Producing Quality Program Evaluation in Education and Using It: The Washington, D.C., Experience

June D. Bland 113

14. Proceedings of the Practitioners' Conference

Charles B. Stalford 121

Participants 127

Introduction

Charles B. Stalford
Testing, Assessment and Evaluation Division
National Institute of Education

This volume is a compendium of practitioners' papers about testing and evaluation issues in the schools.

In 1978, five teachers, five principals, and five school research and evaluation staff people were asked by NIE's Testing, Assessment and Evaluation Division (TAE) to help plan its future research. TAE had been recently created as a result of an NIE-wide reorganization, and the time was right for it to seek advice from a body of school practitioners to get its new research plan off to a good start.

Accordingly, the 15 individuals were each asked to write a paper reflecting *personal* concerns from their experiences with testing and evaluation in schools. They were further to suggest research that could lead to improvement in these functions. The writers then attended a 2-day conference in Washington, D.C., during the summer of 1978. There, they presented their papers, still in draft form, for discussion by the entire group plus selected TAE staff. Following that conference, the papers were revised and submitted to NIE.

Excerpts from 13 papers (one invitee subsequently declined to participate and one paper was jointly authored) and a summary of the conference proceedings are published here. The conference was a lively and stimulating affair in which many new perspectives and insights were generated. In part, this was due to the unique composition of the practitioner group itself; it is not common for teachers, principals, and research and evaluation staff to sit in the same room for an extended period and exchange views on testing and evaluation in which all are intimately involved, but from different and often conflicting perspectives.

For example, teachers may often be asked to cooperate with local evaluations or testing initiatives without having the opportunity to discuss or understand their rationale. Such lack of communication may lead to frustration and minimal cooperation, often at the expense of the evaluation or testing activity. Variants of this phenomenon, substituting principals or research and evaluation directors or other administrators, are easily imaginable.

For TAE staff, the conference proceedings and papers were enlightening. TAE sought through this activity to broaden its perspectives beyond the academic research circles it might often consult in its planning. Participants were accordingly discouraged from writing papers in a traditional academic mode, with literature reviews and the like; rather, they were asked to write about the "heart" of local testing and evaluation matters that concerned them. It was TAE's belief—sub-

stantiated by the project outcomes, we believe—that such insights were absolutely essential to formulate a *useful* plan of research to improve testing and evaluation practices in schools.

Most of the participants were referred to NIE by national professional associations and offices. In particular, the NEA and AFT suggested lists of candidates from which all but one of the teachers were selected, and the National Associations of Elementary and Secondary School Principals did likewise for the principals. With one exception, research and evaluation staff were suggested by Dr. Michael Kean, Executive Director of the Office of Research and Evaluation of the Philadelphia Schools and Chairperson of the Large City Directors of Instructional Research, a school-oriented special interest group loosely affiliated with the American Educational Research Association (AERA). A full listing of participants is found at the end of this publication.

TAE makes no claim that the participants are necessarily representative of all such practitioners in the country nor that the views and conclusions expressed in their papers are necessarily so representative. This activity was planned as a quick response to the need for a new governmental unit to gain assistance in its planning, not as an exhaustive sampling of practitioner views. A quick glance at the diversity as well as positions of the participants does support, however, the idea that their views should be considered seriously.

The activity itself plus subsequent contacts with the participants has, in fact, been practically used by TAE and the Institute. Actions taken by TAE in its research plans which are responsive to conclusions reached in this project are described in the conference proceedings (chapter 14) of this publication. It was our belief that the papers and proceedings summarized here were also of sufficient general interest and potential use to educational policymakers and practitioners that they should be published.

Three further comments are in order before proceeding to the body of this publication. The first describes the mission of NIE's Testing, Assessment and Evaluation Division (TAE) and amplifies the significance of this activity. The second comment suggests a further limitation to consider when reading these papers. The third comment synopsizes the content of the papers and describes procedures used by the editor in abridging the original versions for publication herein.

Mission of the Testing, Assessment and Evaluation Division in NIE

The Testing, Assessment and Evaluation Division seeks to improve the practice of testing and evaluation in schools through research. While the testing and evaluation functions are closely interrelated, TAE has separate thrusts in each. In testing, TAE is concerned both with practical considerations about test construction and use and with more basic research to better link testing with instruction.

Thus, for example, within the testing area, TAE sponsored a series of conferences throughout the country in 1979 to meet teachers' needs for information about effective testing practice. These conferences were one outgrowth of a national conference on testing sponsored by NIE at the initiative of HEW Secretary Joseph Califano in 1978.[1] In addition, it has funded a grant to the American Federation of Teachers (AFT) to develop materials for teachers intended to show how testing can be better linked to instruction. In part, this research is based on an assumption that testing *should* be better linked with instruction to enhance learning, it is also intended to downplay the emphases frequently criticized in traditional testing programs, namely those of student sorting and selection and those serving accountability-related purposes.

Research in testing has been stimulated by an NIE-sponsored conference of researchers in testing, psychology, and related social science disciplines at Falmouth, Massachusetts, in August 1978. Conferees emphasized theoretical linkages between testing and cognitive psychology and suggested numerous areas of research appropriate for further investigation.[2]

A national study of minimum competency testing (MCT) programs was initiated in 1979 and provided preliminary descriptive information on practice in this area. Setting aside further evaluative study of MCT using traditional social science methods, TAE is sponsoring a "clarification hearing" on the topic. This hearing, to be held in 1981, will embody judicially inspired rules of evidence and procedure to allow opposing views about issues in MCT to be presented and discussed. The hearing will be videotaped for distribution throughout the country, together with supplementary written materials, so diverse lay and professional audiences can be helped to understand MCT better. TAE also funds a diverse program of research on test use and design at the Center for the Study of Evaluation (CSE) in the UCLA Graduate School of Education. CSE's research on testing focuses on assessment of writing as well as other basic skills.

Those interested in learning more about research on testing funded by TAE may write for further details to Judith Shoemaker, Team Leader for Testing, National Institute of Education, 1200 19th Street NW., Washington, D.C. 20208.

TAE's research on evaluation is concerned with program evaluations: that is, evaluations of legislatively or administratively inspired educational programs designed to serve specific populations and bring innovation to schools. Such programs include Title I and several other Titles under the Federal Elementary and Secondary Education Act plus other special Federal programs. In addition, they include a wide variety of exemplary programs funded under State and local auspices, as well as demonstrations having national visibility which may be funded periodically through foundations and other sources. Thus, for example, TAE is now funding an evaluation of the Reverend Jesse Jackson's initiative in urban education, PUSH for EXCELLENCE, which is funded from both Federal and non-Federal sources. At the same time, TAE is funding research on ways to increase the use of such evaluations by policymakers and school people alike.

In an area most closely related to the subject of this publication, TAE has funded research on how evaluation functions are performed in local school districts. One part of this has been conducted by CSE. As an outgrowth of that research and with the results of this project, in 1979 TAE expanded research to improve the practice of program evaluation at local and State levels. This research will study exemplary uses of evaluation and testing to improve instruction and will explore ways to provide better technical assistance locally in testing and evaluation.

Those wishing further details about TAE's program of research on evaluation may write Charles B. Stalford, Team Leader for Evaluation, National Institute of Education, 1200 19th Street NW., Washington, D.C. 20208.

Limitation on Scope of This Publication

The second preliminary comment in this introduction notes a limitation on the scope of this publication. A purposely limited sample of practitioners, specifically those with local level responsibilities, was invited to this activity. Representatives of other constituencies and institutions concerned with local testing and evaluation practices were not present.

For example, representatives of State education agencies did not participate in this activity. No slight to State level concerns about testing and evaluation was intended. Among other things, a program of research on evaluation funded by TAE at the Northwest Regional Educational Laboratory in Portland, Oregon, devotes substantial attention to State level interests in evaluation. We in TAE see a great need for better information about State level concerns in testing and evaluation.

Neither did representatives of test publishers participate in this activity. Standardized tests are frequently (although not always) criticized in the papers herein. If a "hearing" had been held on the adequacy of standardized tests per se, appropriate representation of all pertinent views, including those of publishers, would have been sought.

The foregoing discussion is to emphasize to the reader that this publication reflects local practitioner views on testing and evaluation. On occasion, comments critical of Federal as well as State agencies and other matters are found in these papers. NIE does not necessarily endorse conclusions stated herein about the adequacy of various policies or practices as perceived by local practitioners. The papers in this volume are not published to force the reader to "choose sides" between the merits of Federal, State, and local or any other combination of professional and constituent viewpoints on the issues discussed.

The volume *is* published to provide a glimpse of "how it is" for local school people trying to administer and cooperate with testing and evaluation programs and also to air their suggestions about how future research could improve such programs. The publication is intended to be a resource for better understanding, not a policy document.

The reader might well look for generic themes which run throughout this volume rather than concentrating on the specifics in individual papers. Some but not all such themes are the varying degrees of rationality which national testing and evaluation issues assume locally, the need for improvements in communicating useful information about testing and evaluation locally, and the need for more cooperation among the various parties involved in testing and evaluation issues in schools, as well as across different governmental levels.

Editing and Content of Papers

The chapters herein are abridged versions of the papers submitted to NIE. The originals typically were 30 double-spaced pages and hence too lengthy to publish in their collective entirety for a busy audience.

The editor, with the concurrence and review of the authors, has abridged them in the aggregate to roughly half their original length. In individual cases, the abridgements have been more or less.

In editing, certain rules were followed: (a) the basic theme and structure of the original paper were retained; (b) certain extended discussions of theoretical topics—e.g., mastery learning and Rasch scaling procedures—which are in more academic sources have been deleted; (c) where papers overlapped in subject matter, some sections have been deleted; and (d) extended examples illustrating central themes have been deleted.

In every case, a synopsis of deleted materials is found at the beginning of each chapter for the reader's benefit.

The chapters are presented herein in the overall sequence followed at the conference: those by teachers, then principals, then research and evaluation staff. A brief synopsis of the chapters' content follows.

Judy Singleton, a special education teacher, writes about the special problems learning disabled and other "exceptional" children face in testing and evaluation. William Moore, an elementary school teacher in Florida, describes his concern with local implementation of the Florida Statewide Education Accountability Act and Pupil Progression Act. Without slighting Moore's paper a bit, in view of the controversy surrounding minimum competency testing in Florida reflected in the court case *Debra P. v. Turlington*, which occurred since the paper was written, his is particularly one for which a State level response would have been appropriate if TAE's purpose had been to comprehensively "judge" that particular State legislation. This, however, was not the intent. Again, the intent is to provide insights into testing and evaluation issues as seen at the local level.

Staten Island teacher Edward Cypress expresses concerns about the adequacy of standardized testing programs, in particular New York's citywide testing program, and describes some successful steps taken to cope with it.

In the final paper, teachers Myrna Cooper and Maurice Leiter present a comprehensive statement of teacher concerns about testing. Both writers are teachers whose broad view of the subject is in part due to their experience in dealing with teachers in educational and employment matters through their work with the United Federation of Teachers (UFT) in New York City, as well as their classroom background.

Parker Damon, an elementary school principal, describes the extensive and laborious steps he has initiated in his school to avoid standardized testing and, where such testing is necessary, to structure its use in ways meaningful for students, parents, and teachers. Luis Mercado, also an elementary school principal, provides his views about why standardized testing does not meet the needs of children in his school and about what kind of testing programs would meet those needs.

Blas Garza and Claradine Johnson describe principals' frustrations with testing and evaluation procedures employed in reform efforts in their schools. Johnson, until recently a high school principal, focuses on an attempt to improve the total climate in a secondary school, involving social as well as academic factors. This is the only paper in this volume which specifically goes beyond instruction-related testing and evaluation concerns. Garza, an elementary school principal, describes the substantive but sometimes ineffectual role which testing and evaluation information has played in efforts to improve reading in a school concerned with bilingual education. In addition, he describes frustration over overlapping Federal and State requirements for program evaluation and testing, a topic of considerable interest to TAE.

Among the research and evaluation staff, Michael Kean outlines the functions of a big city research and evaluation office and its clienteles. Kean highlights the paradoxes in conflicting demands placed upon such an office and offers a comprehensive agenda for research to improve testing and evaluation locally.

Victor Doherty, Assistant Superintendent of Evaluation in the Portland, Oregon, schools describes a long-standing effort there to improve learning through district-wide program objectives used in conjunction with Rasch scaling procedures as an alternative to traditional norm-referenced testing. The chapter highlights such local needs as systematic support for teachers to develop instruction goals, better teacher education in measurement, and more articulation of textbook adoption procedures with local instructional goals, as well as benefits claimed for Rasch procedures.

Ronald Banks, Director of Evaluation of the Buffalo, New York, schools, focuses on three problems in his large urban district: funds for evaluation, relationships between evaluators and other school staff, and use of evaluation findings in the media.

Freda Holley and June Bland discuss factors related to use of evaluations and describe some of their successful experiences. Holley, Director of Research and Evaluation in Austin, Texas, illustrates both good and bad influences on evaluation with three case studies plus some general observations. In an editorial style indicative of her creative approach in Austin, Holley includes drawings to help communicate her message. Bland, Assistant for Evaluation in the Division of Research and Evaluation in the District of Columbia schools, describes an award-winning model used for evaluation of D.C.'s Title I program. In this case, the evaluation model stressed greater participation by teachers and staff in drawing up evaluation plans than is ordinarily the case.

Following these chapters, the final chapter summarizes the discussion about the papers at the 2-day conference in Washington, D.C. As appropriate, highlights and convergent and divergent areas of thinking in the papers which were revealed at the conference are described. In addition, further steps taken by TAE in research on local testing and evaluation since the conference are identified.

We at NIE are grateful for the efforts of these school people and the insights into local testing and evaluation needs they have conveyed. We hope a larger audience finds them useful through this publication.

Notes

1. For details, see *Achievement Testing and Basic Skills; Conference Proceedings* (Washington, D.C.: National Institute of Education, February 1979).
2. For details, see *Testing, Teaching and Learning: Report of a Conference on Research on Testing* (Washington, D.C.: National Institute of Education, October 1979).

Teachers

1. *Testing Concerns of a Special Educator*

Judith Singleton
Fairfax County, Virginia, public schools

Barry slouched at a cluttered table in a section of the noisy intermediate school guidance office. Guidance is a beehive of activity and Barry had an excellent peripheral seat from which to observe the typical goings-on. The four phones were busy as counselors and teachers conferred with parents, sought information from superiors, or made personal calls. The receptionist's office overflowed with students bringing down the morning attendance reports or sullenly announcing they had been thrown out of class and had to see a counselor. A main office secretary chatted with her guidance counterpart as she ran the copying machine.

Barry reluctantly shifted his gaze from this bustle back to the black and white test form in front of him as he heard the next question being asked—or rather shouted—from an adjoining room. You see, Barry was failing several subjects this year and had been peremptorily summoned from class to take a quick intelligence test so the school could see what was the matter with him. The tester, however, was a very busy man and was sandwiching Barry's intelligence test in among his regular duties, phone calls, and confrontations.

Karen's blue eyes were marred by red as she dejectedly left her English classroom after the unit test on a Dickens novel. This eager student with an IQ of over 120 sobbingly announced she had just failed another exam, even though "I knew it; I knew all the answers. I could have told her but I couldn't get it all down. I needed extra time to finish." She took her third tissue from the box on the resource teacher's desk, wiped her eyes, and went disconsolately on to her next class.

Karen was right, right in everything she had said. She had failed the test even though she probably had known all the answers, even though she had a reading comprehension level 2 to 3 years above grade and above that of many of her classmates, even though she had a passion for Dickens, whom she frequently selected for independent reading. She did need extra time. She needed extra time for the physical act of handwriting, made difficult for her by poor fine motor abilities; extra time for the torture of spelling by sounding out, searching for correct sound/letter correspondences, writing the word, erasing it because it didn't look right, and trying again. She needed extra time to make the complex conversion from inner language to written symbols. You see, Karen is a special student not only because of her above-average intelligence but also because she is a learning disabled student in a regular class.

A section on mainstream testing of exceptional children has been deleted from the full version of this chapter—*Ed.*

These small personal tragedies did occur. I was there and I witnessed them. The first was an isolated incident I saw a few years ago and hope never to see the like of again. I struggled in the second incident to respond somehow to the unhappy girl. But this second incident is not isolated. A variation happens daily to one of my learning disabled charges and it tugs at my heartstrings every time. I strongly suggest that, in both instances, the child is not failing a test. Rather, the testing system is failing the exceptional child. This is the problem statement of the chapter.

Concern for the Child

The two introductory vignettes are presented to dramatize the damage wrought by deficient evaluation programs conceived in haste in the early 1970's as special education programs burgeoned and to illustrate the importance of addressing testing and evaluation issues.

Consider the loss of learning experiences Barry may suffer if he is allowed to languish in regular education without provision being made for his learning disability, or his hearing loss, or his emotional burden, whatever it is that is contributing to his school failure. Consider the injury to Karen's self-esteem as she regularly does her best and fails. Consider the behavioral manifestations that may result from a past of unrelieved frustration and the prospect of an unchanging future.

Studies are pointing to a relationship between juvenile delinquency and learning disabilities. *Today's Education* (Vol. 66, no. 4, p. 42) mentions a Government Accounting Office finding that 26 percent of incarcerated delinquents evaluated in two States had learning disabilities.

Consider the burden of numbers of handicapped students who drain society rather than contribute through productive lives and tax dollars. In the same article, *Today's Education* refers to an estimate that 75 percent of high school students with learning problems leave school not only unemployed but unemployable.

> *Special children have special needs*
> *... in a nonspecial environment.*

Equal with the task of originally diagnosing a learning handicap in order to provide appropriate education is that of frequent reevaluation. Retaining a child in a special program when he or she has been misplaced initially or has progressed to the point of no longer requiring that program is a greater danger in the school system without adequate testing personnel and evaluation safeguards. Consider the numbers of black children whose differing cultural and linguistic backgrounds led, in the past, to their placement in programs for the educable mentally retarded.

The goal of all education is enabling young people to become satisfied, productive citizens, able both to cope with and to enjoy life. Special educators strive to surmount the additional hurdles handicapped children face in achieving this goal. Competent testing and evaluation is the keystone to facilitating rather than impeding learning, enhancing rather than blemishing self-image, nurturing the behaviors of adjustment rather than maladjustment.

A companion to this humanistic impulse to test and evaluate correctly is Federal legislation. Spurred by growing national concern for the rights of all minority groups, since 1965 Congress has legislated support for the educational rights of that minority composed of handicapped children.

The Education for All Handicapped Children Act (1975), better known as PL 94-142, requires all school systems to respond to the needs of these children by

- Providing free and appropriate public education for all handicapped children.
- Developing an individual education program (IEP) for each exceptional child.
- Conducting comprehensive team assessments which are neither racially nor culturally biased.
- Guaranteeing due process for children and parents throughout the evaluation and placement procedure.
- Placing handicapped children in their least restrictive environments that is, in the general education program if the child can be successful there.

To my mind, an adequate local testing and evaluation program must function for each of these five requirements for the program to be meaningfully implemented. Appropriate considerations are:

- Testing and evaluation must be conducted in order to determine what is "appropriate" education.
- A meaningful individual education program can be based only on thorough diagnosis.
- Tests and measures constituting a valid, unbiased, comprehensive assessment must be available.
- Legal consequences can result from misdiagnosis of or failure to diagnose the exceptional child.
- Adjustments can be made to the regular program to insure the success of handicapped learners.

Learning Disabilities

Special children have special needs both as special populations with special labels and as special children working in a nonspecial environment. In the remainder of this chapter, I will detail the testing concerns of the special educator, describe attempted solutions, and offer suggestions for research and development.

I will approach these topics from the point of view of the learning disabilities teacher because my firsthand experience is in this area. Nevertheless, I am confident that my concerns and goals were widespread throughout other special education areas such as mental retardation, emotional disturbance, and vision and hearing impairments. Moreover, my point of view is that of a teacher in a specific system. Other systems no doubt, vary in their support to special education personnel.

A description of the situation in which I work will serve to illuminate the testing problems which I perceive. My students, seventh and eighth graders with learning disabilities,

> are those children who have a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations. Such disorders include such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. Such term does not include children who have learning problems which are primarily the result of visual, hearing, or motor handicaps, of mental retardation, or emotional disturbance, or environmental, cultural, or economic disadvantage. (P.L. 94-142).

As resource students, they are considered able to function adequately in the mainstream with teaching and support in the resource room. They study 5 class hours daily with regular education teachers and 1 class hour daily with me.

To provide these students appropriate education in their least restrictive environments, I endeavor to

- Teach approximately 24 children identified as learning disabled resource students.
- Write and implement IEP's (individual education programs) for these students.
- Complete all other paperwork on these students, such as computer updates, annual reviews, triennial reviews, progress reports, and end-of-the-year reports.
- Confer with each student's five regular education teachers to monitor classroom progress and provide appropriate help in completing assignments and taking tests.
- Serve as chairman and secretary for the local screening committee which supervises evaluation and program changes for all special education students in the school as well as for students referred for evaluation.
- Inventory, select, and order special education materials.
- Conduct testing and prepare evaluation reports on children referred to the local screening committee.
- Confer with parents of students in the program as well as those of students tested.
- Provide inservice training for the regular education faculty.
- Attend, and occasionally present material at special education inservices.
- Perform school duties such as attending faculty meetings and PTA, taking bus and hall duty.

Special Education Testing

Testing within special education is conducted for two purposes: first, to determine if special education placement is warranted and, if so, what program is appropriate; and, second, to substantiate the need for children already enrolled in special education to change to a more or less restrictive program or to continue the ongoing program. My concerns about testing in these areas are also two—the adequacy of the testing program and the length of time required for evaluation and placement.

I am going to use Barry, the boy whose intelligence was measured in the guidance office, to illustrate my concerns. Barry was referred to the local screening committee because he puzzled his history teacher. The boy was attentive to discussions and audiovisual presentations; he had a good fund of general information which he fluently shared with the class. On the other hand, Barry was disruptive during class work periods, handed in few assignments, and failed most of the tests he bothered to take.

For the sake of brevity, let us assume that all work preliminary to testing—reviewing the cumulative folder, interviewing teachers, discussing Barry at the local screening committee meeting, obtaining oral and written permission from Barry's parents for testing, negotiating Barry's release from class for testing—has been accomplished. This stage alone can consume several weeks.

The first step is to determine Barry's intelligence range. Remember, the definition of learning disabilities excludes mental retardation. I immediately have a problem with test adequacy. I have scant confidence in Barry's guidance office performance so I use the only intelligence measure

I have at my disposal. Barry scores in the high average range. However, I have measured only Barry's intelligence in auditory, verbal, and memory areas, the components of intelligence this test measures. I am ignorant of Barry's abilities in the visual and performance areas—those areas in which, from teacher comments, I expect the probable learning disability lies.

To get a measure of intellectual functioning in these areas, I must request that a school psychologist administer a test such as the WISC (Weschler Intelligence Scale for Children). This means another meeting of the local screening committee (which meets at 3- to 4-week intervals), oral and written permission from the parents, and placement on the psychologist's testing list. Since the psychologist or psychometrician assigned to my school spends less than a full day a week with us, Barry's name will move slowly to the head of the line.

Meanwhile, Barry and I pursue a second area: achievement testing. I will compare the level at which Barry is achieving in academic areas with the expectancy level which we have "learned" from the intelligence test. Barry is given three achievement tests: a comprehensive measure of spelling, arithmetic, and reading recognition; a reading comprehension measure; and a diagnostic mathematics test. From this testing I conclude that Barry is achieving below his grade and ability levels in reading comprehension, reading recognition, and spelling. His mathematical skills appear adequate.

Now I conduct a third type of testing. I am searching for an indication of a perceptual disorder which may be handicapping Barry in performing at his ability level in all areas. Again, for brevity's sake, we will assume that Barry's physical, vision, hearing, and speech/language examinations are normal and that records and interviews uncover no environmental, cultural, or economic disadvantage. If it seemed warranted, I would request a psychological evaluation for possible emotional disturbance.

Testing adolescents in the perceptual area is difficult. Tests which purport to measure perception and which are available to me are either standardized on younger populations or not standardized at all. Therefore, they have limited value. I have heard rumors of a recently published perceptual test which has been standardized on adolescents. However, it is not available to me for use at this time.

Nevertheless, I reasonably establish through informal testing that Barry exhibits deficits in visual organization, memory, and sequencing, as I had suspected. Therefore, I submit my data to the local screening committee. This committee recommends placement in a learning disabilities program. At a later date, placement is officially granted by the area eligibility committee, which meets monthly.

Barry's parents are then informed of the decision and may accept or reject the program change. Should they accept, after all of the paperwork is completed and Barry goes on the computer, his schedule will be changed to include appropriate help. Judging by my experience, evaluating and placing Barry would have taken a minimum of 3 months, most likely longer. Meanwhile, Barry has endured 3 or more months of continued frustration in several of his general education classes.

Special education students are reviewed annually for continued placement. Triennially they undergo evaluation identical to that required for initial placement. Let us assume that Barry has now come to one of these points. I find that Barry's achievement scores in his deficit areas of reading comprehension, reading recognition, and spelling have neared his ability level. He is no longer failing any subjects. Ergo, the local screening committee and the area eligibility committee move, under existing guidelines, to dismiss him. Barry does not need special attention any more...maybe.

Maybe Barry has remediated or compensated for his learning disabilities to the point that he will now be able to learn successfully in general education. This happens not infrequently among learning disabled resource students.

Maybe, on the other hand, my sixth sense is accurate. Despite his improvement on test scores, my daily observations tell me that Barry will not be successful. It is true he has demonstrated competence in academic skills. However, these skills were evaluated in isolation, without time requirements, in a comfortable environment; that is, Barry was given individual standardized tests by the resource teacher. Barry was not asked to integrate these skills with others; he was not pressured by time or by fear of failure in front of his peers—he was not asked to perform under classroom conditions. Success on an individually administered reading comprehension test does not insure success in reading and understanding textbooks.

In the classroom, Barry is asked to perform tasks far more complex than those tested. For an example, consider the process of notetaking. Barry has mastered the individual skills involved—listening, remembering, handwriting, to name a few. Notwithstanding, he is unable to fluidly integrate these isolated skills and take adequate notes. Notetaking is more than the sum of its parts. Barry is required to listen to new material while mentally extracting the high points of material previously heard while writing prior material. . . .

Unfortunately, these organizational and sensory-integration deficits are not measured by the standardized tests used to determine continuing eligibility. I grit my teeth as the committee scans the test data, sighs over the evaluation guidelines we all must follow, and recommends Barry's dismissal.

To my way of thinking, Barry has been penalized for showing skill improvement. The child who has received special help for a number of years *without* making significant advances is eligible to continue. I have come to believe that this child's neurological deficits are such that improvement is not likely to occur now that he is well into adolescence. But Barry, the child who was progressing and who shows promise of more progress, is no longer eligible for continued help because of that progress!

In no way do I intend this section as a criticism of P.L. 94-142's provisions nor of my school system's policies and procedures. Both Congress and my system strive to protect the legal and educational rights of children. They are reluctant, and rightfully so, to label any child as "special" until thorough testing and evaluation indicate that special services are needed for that child's appropriate education. Special educators are now laboring with the proverbial one hand tied behind the back. Adequate evaluation instruments and personnel are not provided to satisfy the good intentions.

In addressing these inadequacies, I have personally tried these approaches:

- Lobbying for appropriate evaluation instruments and requesting psychological assessments.
- Including additional school personnel in the evaluation process.
- Reducing the time spent planning for my 24 daily students.

Lobbying—The first approach has borne no fruit as yet. My supervisors share my frustration with existing measures but have not been able to alleviate the problem. Since the number of special education students is increasing and since the increase in the number of pages of paperwork required is in direct proportion, the development of an adequate evaluation system seems increasingly remote. Indeed, in this day of spiraling costs and taxpayer revolt, cuts in staffing are more likely than additions.

Including Additional School Personnel—The second approach has been successful in some degree. This year I have been helped by a cooperative reading teacher who has undertaken the reading evaluation of students referred to the local screening committee. This is voluntary on her part and is in addition to her own workload. The school nurse has taken it upon herself to produce the required vision and hearing data. The speech clinician has been prompt in submitting her evaluations. Nevertheless, these procedures are not institutionalized. Because I have no authority to require help, the help I receive is given to me on a personal basis only. Furthermore, the bulk of the evaluation must be accomplished by a person with knowledge of testing and experience in special education. In my school, that's me.

> *Planning is left to the 15 minutes before school and the 45 minutes after.*

Reducing Planning Time—I am quite uncomfortable with the third approach although it is successful in that I do create more time for testing. The tradeoff is that I have less time for everything else. It is not unusual for me to teach four periods and test the remaining two periods. Planning is left to the 15 minutes before school and the 45 minutes after—if no conferences are scheduled, no faculty meetings are held, no phone calls must be made or received. This exhausting schedule, broken by a half-hour lunch period, takes its toll on me as a teacher. The effects on my students are worse. They are deprived of the quality of help I should and could give them.

Suggested Solutions

Remedies for the problems educators confront in testing and evaluating the exceptional child are not in themselves exceptional. However, exceptional boldness and commitment will be required in their application if, as so often happens in the educational bureaucracy, the cure is not to become worse than the disease. I offer five suggestions for addressing the issue:

- Decrease in the student/teacher ratio.
- Education of teachers in the testing needs of special populations and in testing skills.
- Supervision in the testing area by administrators.
- A full-time testing consultant in the schools.
- Appropriate evaluation instruments.

Decrease in Student/Teacher Ratio—Individualized instruction cannot occur en masse nor can individual attention be given in a crowd. In theory, the learning disabilities program is one in which children receive an IEP tailored to their specific needs and individualized instruction in their areas of deficit. In truth, my resource classes average six students; class periods are 55 minutes from bell to bell. If the entire period were devoted to this advertised individualization, each of my students would receive 9 minutes and 10 seconds of individual help per day. If I were to allow 5 minutes of planning time to assessing each child's daily work and providing new assignments, I would use 2 hours per day in this way.

A conscientious classroom teacher has similar problems. It is far from unusual for him or her to deal with 150 students per day. How much individual attention, assessment, and adaptation can be reasonably expected?

I have read a suggestion for counting special education students differently in computing the student/teacher ratio. Children with varying handicaps are counted as 1-1/2 or 2 or more units depending upon their degree of disability. Thus, a teacher of 20 regular and 5 learning or otherwise disabled students would be considered to be teaching 30. The school would maintain a 30:1 ratio while the teacher would be better able to integrate the handicapped learners into the mainstream.

Decreasing student/teacher ratios in both general and special education is a major hope for meeting the educational needs of our young people.

Teacher Education—I was allowed to graduate from a reputable college with a secondary teaching certificate having had 18 hours of education credits—approximately 20 minutes of which dealt with special education. I never heard at all about learning disabilities, which had only recently been "discovered" at major universities and not yet made known in the hinterland. I never had course-work in testing anybody.

Aware of such deficiencies in teacher education, many States are increasing requirements for certification and recertification. I applaud the efforts but criticize the courses. Grade inflation and snap courses are not limited to undergraduate days. In my graduate statistics course, most of the students did not know how to square a number. In my graduate tests and measurements class, that 100 was the standard IQ score came as a revelation to many. Two class sessions of 15 in my graduate course on exceptional children were used by the instructor to show medical films on reconstructive plastic surgery. A starting point in improving our children's educations is improving that of our teachers!

Supervision—Administrators wear many hats, as do teachers. I wonder if there is one labeled "teacher adviser" as well as one labeled "bus duty," or "paper pushing." In 6 years of teaching in four different school systems, I have been observed three times by two principals, once by a department chairman. In two of the systems I was never observed. I have been required to communicate lesson plans only when I was absent. Testing and evaluation procedures have never been supervised.

In an ideal school, an administrator would serve as a head teacher, observing, guiding, and assisting teachers in attaining proficiency in their profession. Administrators should have a responsibility to look over teachers' tests as well as their lesson plans, noting the strong points and suggesting corrections for the weak. We educators routinely assume that the student has failed a test. It may well be that the teacher has failed to provide an adequate measuring device. The teacher is not called to account for his failure; the child is. Enough poorly designed tests have been thrust into my hands by frustrated Karens and Barrys for me to know that teachers do indeed fail and do need guidance.

Testing Consultant—A full-time testing consultant in each school is the approach to the testing and evaluation issue most likely to be successful. I envision this person assuming a multifaceted role which would serve administrators, special teachers, regular education teachers, students, and parents. Here is my conception of the testing consultant's responsibilities:

- Provision of testing and evaluation for students referred to the local screening committee.
- Presentation of inservice training to teachers on test construction and administration.
- Aid to individual teachers in developing satisfactory testing skills.
- Administration of regular tests to special students when modification such as reading aloud is necessary.
- Teaching of test strategy for both standardized and regular classroom tests to students.

- Conferences with students and parents about various tests that have been or will be taken.
- Coordination of the school's standardized and minimum competency testing.

I realize I have outlined an ambitious program. A qualified testing consultant would need a regular class teaching background, experience in educating the exceptional child, knowledge of tests and measurements, strong interpersonal skills, administrative ability and...clout.

Authority is the primary quality. An error which has impeded the integration of special education with general is the placement of special educators on a par with regular ones. *Given the hostility and inertia the special educator faces every day, he or she needs firm support from above to achieve assigned goals*. The testing consultant therefore needs to be perceived as neither a special nor a regular educator but as a facilitator for both.

Appropriate Evaluation Measures—Suggestions in this area were contained earlier in the chapter as the testing procedure was reviewed. *To reiterate, perceptual tests standardized on adolescents are not yet in wide use*. Nor are there accepted evaluations of the organizational and intersensory processes so critical to the learning experience. Although there are dangers inherent in subjective analysis, this type of review by competent educators should be more respected in planning for the exceptional child.

Areas for Research and Development

Exceptional children like Karen and Barry are the focus of a contemporary growth industry, similar to computers in the sixties and solar energy today. Note these figures: in 1973, the administrative district in which I currently teach had no established program for learning disabled adolescents. In 1974, the year in which I was hired, nine teachers began the secondary program. The school year 1977-78 saw 34 teachers working in a variety of programs for the learning disabled adolescent. This is a growth of 400 percent in 4 years. The number of teachers will increase by 25 percent in the 1978-79 school year. Growth in any industry at such a furious pace is fraught with pitfalls. In special education, the existing testing system is a deep and dark one, as we have seen.

Issues in testing and evaluating the exceptional child which research and development might valuably pursue have surfaced throughout this paper. These are restated and others mentioned below.

- Investigation of existing testing and evaluation programs in the Nation's school systems with the compilation and dissemination of useful information.
- Development and/or funding of pilot programs for testing consultants.
- Programs for upgrading teacher competence in testing and evaluation.
- Research on the effect of lowering the student/teacher ratio on successful mainstreaming.
- Investigation and sponsorship of alternative testing and evaluation methods to the "paper and pencil" norm.
- Development of comprehensive, unbiased testing instruments.
- Consideration of if, and how, the exceptional child will cope with minimum competency standards.

Research and development agencies have potential in enabling school systems to erase the barriers between special and regular education, thus easing the special child into the mainstream. This will help all educators work to fulfill not only the letter of law but also the spirit: assuring our children appropriate educations and opportunities for optimal personal development. After all, aren't *all* children special?

2. *Florida's Standardized Testing Program: A Tool or Weapon?*

William Moore
Sanford, Florida, public schools

In 1976 the Florida Legislature passed the Educational Accountability Act. Since the passage of this act, student assessment and the testing of students has become the major topic of conversation in the educational community, the press, and many households in Florida.

The main intent of this legislation was to establish a system of educational accountability by providing and guaranteeing that certain worthy things be done by the State, district, or school. The legislation is centered on two main provisions: that the State develop minimum performance standards and that a statewide assessment testing program be administered.

Along with the Educational Accountability Act the legislature passed the Pupil Progression Act, that includes the following provision: each district school board is to establish a program for pupil progression based upon each pupil's mastery of the minimum performance standards approved by the State. Particular emphasis is to be placed upon the pupil's mastery of basic skills before he/she is promoted from the 3rd, 5th, 8th, and 11th grades. These are the same grades in which the Educational Accountability Act mandates that a uniform statewide testing program be administered to determine the educational status, progress, and the degree of achievement of approved minimum performance standards. The last provision of the Pupil Progression Act mandates that "each district school board shall establish standards for graduation from its secondary schools. Such standards shall include, but not be limited to, mastery of the basic skills and satisfactory performance in functional literacy as determined by the State Board of Education."

The effect of this law is to require that every public school pupil pass a "functional literacy" test prepared by the State Department of Education before he/she can receive a regular high school diploma.

It is the purpose of this chapter to look at and discuss the ramifications that these laws will have on testing and the effect testing, in accordance with these laws, will have on students.

A section on secondary level testing and illustrative details have been deleted from the full version of this chapter.—*Ed.*

The Role of Testing

Functional Literacy Testing

With this greater awareness and emphasis on testing, many questions and concerns about the capabilities of tests and their role in our public education system have surfaced. The focal point of most public discussions on testing so far has been the 11th grade functional literacy test. Vast disagreement has surfaced over what can be expected from a test or a series of tests.

According to the Florida State Department of Education, functional literacy is "the ability to apply basic skills in reading, writing, and arithmetic to problems and tasks of a practical nature as encountered in everyday life." This definition seems to be agreeable to most people as long as community expectations are kept in mind. Using this definition, the question arises, Can a paper and pencil test that relies on multiple choice answers truly determine whether or not a person can function in his/her community?

A panel sponsored by the Florida Teaching Profession/National Education Association recently held public hearings throughout Florida. At these hearings several people testified that they knew of students who had failed the functional literacy test but were working successfully and functioning outside of school in a more than satisfactory manner. The report stated, "Examples were cited by parents and teachers where test scores misclassified students as functionally illiterate when there was other evidence to show that the students were both competent in their studies and performing well in their part-time jobs."

Evidence such as this prompts several questions. Is a score on a single test an acceptable basis for making important educational decisions about a student? Can a test be constructed in such a way that it can predict with a high degree of validity that a person who scores low on the test will be unable to function in his/her community?

The assumption by the Florida Legislature is of course that a test by itself can determine a necessary quantity of functional literacy. Joining the legislators in this belief have been the department of education and a large segment of the public. On the other side of the debate are the organized teaching profession, minority groups, and various individuals.

The proponents of the testing program present such arguments as:

1. A high school diploma should mean something more than that a student showed up for classes for 12 years.
2. Through test scores the public can now see clearly how their child's school compares with other schools in their district or State.
3. The public, school administrators, and State legislators can now compare one school district with another.
4. Students must now assume more responsibility for their education.
5. While standards set by the functional literacy test are low, at least there are now some definite standards.
6. This is the only way to stop that horrid practice of social promotion.

Those who oppose the testing program counter with such statements as:

1. The test is culturally biased.
2. The test doesn't test what is being taught in the schools.

3. The system for grading the test is unfair.
4. The test focuses the emphasis of public education on only meeting the minimum standards.
5. So much weight should not be given to any one test.
6. We are fooling the public by leading them to think that testing is such an exact science that it can determine whether their sons or daughters are educated and therefore prepared to meet the world.

Regardless of which side a person takes in this debate, he/she must agree that accepting the results of a single test to make such an important decision is a sharp change from the practice of the past. Common practice has been to consider grades given by teachers who have spent many hours with the student collecting and assessing the different data that are available to the teacher as a basis for assessing functional literacy. Included in this data would be performance on standardized tests and teacher-made tests, and the teacher's own professional judgment.

Before this trend of using only test results for making decisions about students is expanded, much more research should be done on the predictive validity of testing instruments. To continue using a test without empirical evidence that there is a high correlation between failure on such a test and the inability to function in the community may make testing a weapon that is used against students instead of a tool to help them.

Statewide Testing

What has developed as possibly the most integral part of the Educational Accountability Act is the section on statewide testing. According to Florida Statute 229.57(1), "The primary purpose of the statewide testing program is to provide information needed for State-level decisions." As with the functional literacy test, many people believe this provision places an extreme amount of importance on testing. With the passage of this law, it is evident that the legislature made the following assumptions about testing:

a. That by testing students with a uniform test, information can be gathered that is needed for State-level discussions.
b. That testing will assess how well districts and schools are meeting State goals and minimum performance standards.
c. That testing will identify educational needs at the State, district, and school levels.
d. That testing will provide a basis for the comparison of schools, districts, and the State with the Nation.
e. That testing will provide for the improvement of the operation and management of the schools.
f. That standardized testing is a better way to diagnose and assess students than teacher judgment and teacher-made tests.
g. That once a student's deficiencies are determined by a test, they can be remedied by the public school system.

These assumptions by the legislature have stirred debate and resentment within the educational community. Both the debate and resentment center on two major issues: first, can better educational decisions be made by almost solely using test results rather than by relying on professional judgment?, and second, are the time and money spent on testing and compiling results interfering with the teaching process?

With this emphasis on objective test results, it is obvious that the public, through its legislators, wants easily understood criteria for the basis of decisionmaking. It also is demanding hard, concrete facts to be used when decisions are made and not intangible, subjective opinions.

While many critics of the trend towards more emphasis on testing acknowledge the public's desires, they maintain that for the results gained by testing to be meaningful they must only be used in the narrow context for which the test was designed.

The science of testing has not reached the level of sophistication at which broad-based decisions can be made primarily on test results. This seems to be especially true when making comparisons. An example of this in Florida has been the use of statewide basic skills test results to compare two or more school districts to determine which district(s) are doing the best job. Critics have asked whether test results should be used to compare two school districts to determine the district's effectiveness without having identical students in each of the districts.

The question of *how* tests should be used is of great importance. Many teachers are disturbed by what appears to be a trend towards using tests to compare and eliminate students rather than to diagnose and help them. For example, should a test be used to compare two third graders of different intellectual ability, different maturity levels, and different socioeconomic backgrounds? Critics also question what is to be gained by making such comparisons.

> *The question of how tests should be used is of great importance.*

The desire by the public for accountability that can be measured objectively and the realization that there are many limitations on the use of tests have increased the need for research in this area. It is imperative that the experts, in their zeal to justify the growth and funding of research and evaluation departments at the district and State levels, not promote testing as the cure-all for our educational system. It is time, instead, for the experts in testing and evaluation to accept some responsibility for educating the public as to the limitations of written tests.

Districtwide Testing

While the functional literacy and statewide basic skills tests have stirred the most debate statewide, many local school districts have adopted an additional testing policy which has affected the elementary classroom teacher to a greater extent. Section (1) of the Pupil Progression Act mandates that "each district school board shall establish a comprehensive program for pupil progression which shall be based upon an evaluation of each pupil's performance, including how well he masters the minimum performance standards approved by the State board." In one district's (District A) pupil progression plan, the traditional grades of 1 through 5 are broken into 20 levels. For a student to be promoted from one level to the next, he/she must pass the appropriate "level test."

A "level test" is actually a battery of tests. Each test consists of several questions that evaluate the student's performance on a particular standard. A student should master the skills for each level in a 9-week period. If a student fails to "pass" the level test, the teacher is responsible for helping

him/her improve in the areas where he/she did poorly and then for retesting him/her. Theoretically, a student should be tested whenever he/she is believed to have mastered the performance standards for a particular level, thereby creating a continuous progression based on the mastery of predetermined criteria.

While this seems to be a logical procedure for moving students through the educational system, some teachers and parents have expressed serious concerns. They complain that with schools using different textbooks, the districtwide leveling tests do not necessarily test skills in the same sequence in which they are covered in the text. Therefore, they maintain, if the curriculum is going to be constructed around level tests, the district should adopt a textbook series. The level tests could then be based on the textbooks. As it is now, teachers are required to skip continuously about in the textbook, which can be very confusing to young students. Some skills that are covered in the level test may not even be covered in the textbook. While it is recognized that the teachers may still be able to cover such skills with supplementary materials, it is also recognized that this adds an extra burden to an already overburdened teacher and may cut down on the amount of reinforcement material the student has to work with. Therefore the student may not retain the skill. Adopting one textbook series districtwide, they feel, would facilitate full and proper use of the textbook and the teachers', time. It has also been suggested that the professionally written tests that go with the textbook would be a better instrument for leveling students than the district-developed tests.

There are also teachers and parents who maintain it is not always necessary for a student to show mastery of at least 70 percent of previously taught skills before new skills are introduced and taught.

If the student, by policy, must demonstrate mastery of a skill before he can move on, then 6 weeks of his educational life could be wasted. While this leveling program was introduced to provide diagnostic evaluation for individual students, it has actually reduced the amount of individualized instruction that most students receive. The goal for all students now is to pass the level tests. Many students reach this goal with very little, if any, effort, and they are promoted from level to level while not turning in any work or participating in class activities. These students are not receiving the challenging type of education they need.

At the other end of the spectrum, the slow students are being pushed faster than they should be by teachers who are receiving pressure from the administration to make sure their students pass the "required" number of tests. If the students do not progress on schedule, it is a reflection on the teacher. Therefore, there is a tendency to teach the test with little concern that the student retain the skills tested.

The construction of the tests has been another area of concern. The tests were constructed by a committee of teachers and administrators which based the test items upon performance standards that were written by a previously established committee. There has been criticism concerning the appropriateness of some of the performance standards, thereby indicating that there should be more training in the writing of such standards. However, the majority of criticism has been about the test items. Considering the vast amount of revision that the level tests have undergone during their 3 or 4 years of use and the amount of dissatisfaction that is still expressed, the greatest need in training is in the area of writing acceptable test items from performance standards. One teacher, responding to a question about the quality of the level tests said, "These tests should have been professionally produced and field-tested before the teachers ever saw them. Get out all of the 'bugs' before we use children as guinea pigs."

The lack of field-testing of level tests before their actual use, as stated above, is also a concern. Most of the revision of test items has been done after a full year of use. This means that many students may have been leveled based upon poorly written test items and not on their ability.

Teachers and parents have also complained that the curriculum has been altered and limited because so much emphasis is being placed on minimum standards. Because of time restraints, this emphasis on basic skills and minimum standards has brought about the elimination of many creative and "fun" activities that teachers feel are important to the child's educational experiences. One teacher asks, "When does the creative teacher have time to develop an exciting program for her room when all of her time is spent on level tests?"

The demand for accountability through standardization and comparing test results can be very damaging to teacher enthusiasm and morale. "I do not feel that I am permitted to make the professional decisions for my students that I was trained to make. I feel very restricted and frustrated about this. Why get a Master of Education degree in teaching if you are forced to follow such a 'lock-step' program with testing at each step?" is one way a teacher expressed her feelings. Another said, "In education today we are constantly needing to prove our accountability. It's almost as if we are not supposed to have a high opinion of ourselves. We almost dissect the pupils in proving that we are working. I used to be more anxious to get to school than I now am. The part that really distresses me is that I wouldn't want my loved ones to get into education, yet I want good teachers for them."

Society cannot afford to squander a teacher's time on clerical duties.

The appropriateness of teaching to the test also becomes involved in this discussion. Many teachers point out that whenever a lot of emphasis is put on test results and those test results are used to judge and compare, then eventually one of the items that will be judged and compared is the teacher. Teachers also point out that when anyone, including teachers, is evaluated in such an unfair way, he/she will, for survival purposes, adjust to protect him/herself. One way to do this is to teach the test. After all, if the administration and the public are so greatly concerned about students' ability to perform well on tests, then this is what students should be taught.

The most persistent concern expressed by teachers has been how time-consuming level testing is. The concern deals with the relationship between instruction and evaluation. Teachers recognize that both are important and that they are related. In discussing this relationship some teachers have used the analogy of instruction being the dog and evaluation the tail. Using this analogy they feel that in District A's plan, the tail (testing) is wagging the dog (instruction). They point out that with all the testing, retesting, and recordkeeping required with classes of 30 or more students, instruction and preparation time are severly limited. This, according to some teachers, has become a "Catch 22." The teacher is responsible for what the student learns or does not learn, and is held accountable through tests that take so much time that there is very little time left to teach. If money is not available to hire clerks to handle the records that are generated by such a testing program, then the program should be altered to get the most out of the teachers in the area for which they are paid, teaching. Society cannot afford to squander a teacher's time on clerical duties.

Conclusions

When test results are used to make decisions about students, schools, districts, and the State educational system, rigid cutoff scores are established to facilitate the decisionmaking process. The most commonly used cutoff point is between passing and failing. With the functional literacy test, the basic skills tests, and the level tests, there was only one cutoff point. Seventy percent was determined to be the difference between passing and failing. How the score of 70 percent was derived is very vague to many teachers; but how the cutoff point was derived is not nearly as important as its effect on management decisions.

Most decisions in education revolve around the available resources and the establishment of priorities. With the emphasis on meeting minimum standards and on remediation for students who score below the cutoff point, it is only natural for this to become the number one priority and therefore receive additional resources. If there is not a corresponding increase in overall resources, then other sections of the educational system will naturally suffer. In some schools and districts this has become obvious as class size goes up and money for supplies goes down in classes for average and above students.

The comment most often made by teachers in favor of the testing program in Florida is that *now* more of the responsibility for learning is placed on the student. These teachers feel that there is now more of a built-in motivation for students. In their opinion, the ultimate motivation is the functional literacy test. If students realize that they cannot receive a high school diploma without learning the basics, they will become more serious students in the early grades. Others believe that such punishment will be meaningless to students in the lower grades where the foundations are laid. Research needs to be done in this area of testing and motivation.

This entire new emphasis on testing and student assessment developed out of the public's desire for accountability. With the public's loss of faith in their political leaders, the desire for measurable results from government institutions has increased. Since the educational system is the State's largest business, it has shouldered the brunt of the public's dissatisfaction. The political leaders seem more than willing to use teachers as scapegoats. The combination of these attitudes has fostered a loss of faith in the professional judgment of teachers. Teachers are asking why they spent so much time, effort, and money to be trained if they are not trusted as professionals.

It is possible that well-trained teachers and administrators can direct the educational system more successfully if they are left with some flexibility to meet local needs. Should a school in a major city with a student population containing a large percentage of Spanish-speaking students be guided by the decisions made in a school located in a high socioeconomic suburban area? Should the two schools even establish the same minimum standards? Or should they each be managed by trained professionals making decisions on local conditions and needs and not strictly by test results?

Our country is based upon the worth of the individual and the recognition that each person differs from his neighbor. This is especially true with children, who mature at different speeds. The basic modern philosophy of our system of public education has been to work with individual needs of the students. Mass uniform testing does not lend itself to this type of philosophy. When you say all third graders must pass a certain test *now*, individualized differences are not being considered.

We must teach students to do their own thinking so they will be able to adjust to a changing world. If we only train students to memorize knowledge and emphasize only basic skills, they may have great difficulty adjusting as technology mandates. It is impossible today to determine what a

15-year-old student will need to know when he/she is 35. Skills and knowledge that are important now may be of little use in 20 years.

Teachers throughout Florida have what could be called horror stories about the effect of uniform testing on some students. They talk about the emotional stress some students go through, especially in the lower grades. These students often are tested in unfamiliar surroundings by unfamiliar people who demand that they perform tasks that they cannot do. The reasons the students cannot do the tasks vary—fright, poorly explained instructions, poor testing conditions, poorly written test items, lack of academic readiness—all of which cause frustrations that at times can even bring tears to their eyes.

This charge, that our children are being frustrated by this emphasis on testing, has serious implications. If it is true, then the frustrations that the children experience in these instances may adversely affect their relationship with their teacher. How can a young child trust someone that forces him/her into such a situation and then refuses to answer questions or help him or her in any way? Even students in the ninth grade have exhibited extreme anxiety and frustration when they are asked to answer questions based on a reading selection with vocabulary words that they do not understand. This frustration is increased when they ask the test proctor, a normally friendly teacher, and the teacher responds with "I am sorry, but I can't help you." This type of experience will reinforce any negative feeling a student has towards school and can be devastating to a student in the primary grades.

The purpose of the laws which initiated the testing programs was to cause changes that would improve the quality of the public educational system. If the changes that are caused by the testing programs reduce the quality of the public educational system or only bring about changes on paper, then the testing programs should be radically altered or possibly the laws should be repealed. Many teachers, worried about the changes that have already occurred in their classrooms, advocate repealing the laws and stopping what they perceive as the "bastardization" of testing.

This struggle to determine the role of testing in our public educational system is of great importance. Until our political leaders, research and evaluation experts, and the teaching profession can agree as to the proper use of tests and their results, it is recommended that a moratorium be put on the practice of standardized testing. I recognize that this would be a drastic action, but it may be the only way to protect the innocent, both students and teachers. Such an extreme course might force the divergent groups to work together.

Even while this struggle to determine the role of testing in our public educational system is continuing, experts in this field should not only carry forward their research, but, equally importantly, they should also attempt to educate our political leaders and the public about the promises and limitations of testing.

In March 1979, a U.S. District Court in Florida enjoined the State of Florida from requiring passage of a functional literacy test for graduation for 4 years. In its decision, the court did not find that the Florida test was racially or ethnically biased but did find that inadequate notice provided prior to invocation of the diploma sanction was a violation of due process. The court also ruled that the use of the tests for remediation purposes was constitutionally permissible. The case is under appeal—*Ed.*

3. *Making Reading Achievement Tests Work for the Inner-City Student*

Edward J. Cypress
New York City public schools

In this chapter, the writer, as teacher-in-charge of a second grade minischool in New York City, observes the difficulties faced by inner-city children who are compelled to take standardized reading achievement tests.

The manner in which America's middle class has viewed the progress of its minority population has set definite and serious limitations on the latter's upward social and scholastic mobility. Crippling labels denoting inferior mental attributes have been assigned to our Nation's black population. Many more black and Spanish-speaking Americans are classified as mentally retarded than would be expected from their relatively small segment of the total population.

Society's reluctance to accept the socially and economically disadvantaged has always influenced the outcome of devices which have been used to measure achievement, aptitude, and intelligence. The early use of the Binet test classified approximately 80 percent of the immigrants who passed through Ellis Island in 1912 as being "feeble minded." Since the arrivals from southern and eastern European nations were deemed less socially acceptable than were earlier immigrants, their scores reflected this bias.

In an attempt to gain status in areas governed by academic inquiry, psychologists and educators have for the past 65 years turned to science and technology in an effort to develop statistical and scientific methods of evaluation. In doing so, they have devised instruments which for the most part measure the results of cultural deprivation and not scholastic achievement.

With this in mind, I would like to address the deficiencies that are exhibited in current city testing programs based on standardized reading achievement tests and to suggest ways in which these deficiencies can be improved.

A section describing advantages seen by the author in the "Language Experience Approach-Cloze" technique of testing has been deleted from the full version of this chapter–*Ed.*

Limitations of Norm-Referenced Achievement Tests

Current tests which measure reading achievement are of questionable value in educating the inner-city child. Since "the purpose of measurement is to provide information which can be used in improving instruction,"[1] it is imperative that educators take a hard look at these devices to determine if they accurately predict those skills and abilities that are necessary in the development of the student.

Formalized testing techniques are necessary because reading development is a continuous process and must be monitored to determine the effectiveness of the instructional procedures that are being used. However, the limitations of standardized reading achievement tests as tools of evaluation prevent the educator from incorporating test results into the curriculum and making sound administrative decisions based on the results.

Publishers of systems that attempt to evaluate scholastic achievement first look to the total population to determine their norming group. Through an elaborate selection process, a fair sample representing a cross-section of our Nation's youth is chosen as the base against which comparisons are to be made. Most current norming groups purportedly include a representative portion of inner-city children, but they are still not sound foundations against which this group can be compared. Norm-referenced tests, as they are currently constructed, measure narrow ranges of skills and knowledge and do not evaluate the disadvantaged child's potential for achievement.

Inner-city children show a readiness for learning at an early age, but because of their cultural backgrounds and environment, for different sets of materials than their more affluent peers. Socio-economic conditions have instilled in such children a strong desire for survival and independence. However, they cannot relate to the values that the middle class has set for them.

Although the inner-city child must one day function in the larger society, it is necessary to determine his/her level of achievement in relationship to other children of similar backgrounds and experiences. In addition to the norming tables that are currently in use, it is imperative that publishers of educational testing materials, with the help of local educators, set up local norming groups as well. The information provided by comparing disadvantaged children to their socio-economic peer group would be helpful in understanding students as they relate to their subcultures. To do this, it would be necessary to determine the most recent performance of local students.

Publishers, who are naturally concerned with their corporate image, have been reluctant to develop separate norming data. They assert that attempts to do this would be resisted by leaders of the minority community who would proclaim that their endeavors are fraught with racism. If this issue was examined more carefully, economics would appear to be the underlying cause for the unwillingness of publishers to give this matter serious thought. It might cost a great deal of money to construct tests that would be used by a limited market.

At one time there was a strong consensus of what the goals for educating our youth should be. Throughout our Nation, curricula were also similar. Today, both the objectives of education and curricula vary with the needs of the students; but there is little change in the items that achievement tests cover. Thus, the process of norming may not permit the matching of test content with the actual curriculum that is taught to the student.

Inner-city children, who generally have a weak oral language base, begin their education with a poorly developed vocabulary. The sight vocabulary that they acquire generally comes from a

basal reading program and from teacher-directed independent reading lessons. The vocabulary skills that were examined by the Metropolitan Reading Achievement Test, given to New York City elementary school pupils in 1973-74, went far beyond what was taught by the basal readers and embodied a variety of words that had been gleaned from supplementary reading materials. When a switch was made in 1975 to the Stanford Reading Achievement Test, the total scores in reading rose. A review of the test quickly revealed that a word study subtest was included, thus indicating that inner-city children do best in what is school-related (i.e., phonics). They did most poorly in listening-based vocabulary skills that called for a language experience background that they did not possess.

It is interesting to note that the publishers of the Stanford Achievement Tests chose to delete the listening vocabulary subtest in their 1976 edition. Their reason—the time to administer the test extended beyond reasonable limits and children in the early primary grades would be frustrated by the length of the examination.

Technical imperfections in the development of test items, coupled with ambiguous tasks, frequently cause much confusion. This was unfortunately highlighted in a recent standardized reading achievement test that was given to second and third grade students who attend the public school system in New York City.

Review of the 1978 New York City Reading Test Program

The 1978 New York City Reading Test, published by CTB/McGraw-Hill, Inc., was the newly developed, secured version of the California Achievement Test Form D. The examination given to the majority of the students in grades four to nine went smoothly. But second and third graders who were given a lower level of the test found it to be very disconcerting.

The test was divided into four subtests which consisted of phonic analysis, structural analysis, reading vocabulary, and reading comprehension. It lasted 68 minutes. The portion devoted to structural analysis covered the formation of compound words, the selection of two words that form a contraction, and the use of base words and affixes. The tester read instructions to her students before each part and directed them to perform separate and distinct tasks. Within a few moments, it was obvious to many teachers who administered the test that this exercise was becoming disordered and that their charges were frustrated and confused.

The situation was further exacerbated by the inclusion of test items in both a vertical and horizontal fashion. In addition, prior to the last two segments of this subtest, the teacher's manual advised the tester to devote a few moments to the definitions of contractions, base words, prefixes, and suffixes.

Generally, reading programs geared to second and third graders emphasize questions that call for explicit answers. Although many inner-city students can solve inference questions on an oral level, their lack of a developed sight vocabulary and of experiences in solving such tasks prevent them from responding correctly. The subtest of reading comprehension dealt with exercises that tried to determine the students' knowledge of figurative language and their level of critical, literal, and interpretive comprehension skills. More than half of the questions stressed responses that were inferential in nature.

The Board of Education did attempt to anticipate the difficulties that its students were going to face when they took the standardized reading achievement tests. A list of domain groupings was

issued for its elementary school population. Local school districts were encouraged to use this information in the development of test sophistication programs.

To show what could be done to help children prepare for tests, the domain groupings for the Spring 1978 citywide reading tests were used in the following ways in my school:

- A balanced reading program was stressed from the beginning of the school year. I directed the second grade teachers to emphasize decoding skills as well as basic comprehension skills. Through directed and independent reading lessons, teachers were able to see their students' oral and written vocabularies growing.
- During a series of grade workshops, my colleagues and I constructed developmental lessons, which implemented the suggestions made by our district reading supervisor.
- To enable students to determine the implicit meaning of a passage, they were given help in locating certain key words to use as cues or signals to the fact they would be expected to make an inference. The pupils were taught that implicit meanings could be obtained from stated details.
- Stories containing idiomatic phrases were presented to the students, who attempted to discern the meaning of these expressions through the use of contextual clues. Idiomatic language is truly a product of one's culture and should not appear on a reading achievement test. However, without a working knowledge of these expressions, communication and comprehension are limited. Thus, inner-city pupils must become familiar with figurative language and should be able to understand the hidden meanings of such phrases as "make a pig of yourself," "nosebody," "chewing the fat," "chip off the old block," and "putting your best foot forward."
- To augment the vocabulary skills of our students, it was deemed necessary for them to understand the correct use of a word that has multiple meanings. It was suggested that teachers review using contextual clues to determine correct usage.
- A review of compound words was constructed by using graded word lists found in basal readers and in other reading materials. It was suggested that teachers emphasize those words in which both parts give a clue to their meaning. In addition, lessons stressing the joining of words to make compound words were created. This technique directed the student to read a word at the beginning of a line. He/she would then be instructed to read the four words following it and to determine the word that could be added to the first word to make it a new word.
- A game was created to help teachers review the use of contractions. A stack of fifty index cards was prepared so that one half contained contractions and the other half held corresponding verb phrases. Each of 10 participants then received five cards. Each student then took his/her turn and selected one card from the hand of the player who sat on his/her right side. Whenever a student discovered that he/she had a pair (a contraction and a corresponding verb phrase), he/she placed them in front of him/her. The first player who displayed all of his/her cards in this fashion was the winner.

Fortunately, these efforts proved to be successful and the median total grade equivalent score for our second graders was 2.9. This was slightly above grade level. Great efforts had been made to help our students become "test wise."

The results of the 1978 New York City Reading Tests arrived during May and were promptly recorded on the students' cumulative record cards. Since the term was drawing to a close, classroom

teachers paid little attention to the types of mistakes that were made or to the subtest scores that each pupil received. The total score and the teacher's estimation of the child's scholastic ability were, however, factors used in determining class placement for the 1978-79 school year.

Edumetric Tests

I assert that a testing program must consider instructional decisions as its primary goal. Edumetric or criterion-referenced tests measure the scholastic progress of the individual. An evaluation of a student's performance on this type of examination would alert the teacher to the subskills that need to be taught and would determine whether the child has mastered a skill and can proceed to the next level. This form of measurement would also indicate if the student is progressing at a rate comparable with his/her ability.

Through the creation of detailed domain groupings which would touch upon all aspects and levels of reading ability, criterion-referenced tests could be constructed to facilitate a diagnostic-prescriptive approach to reading. This method would set specific instructional objectives that the student should meet. His/her responses would be measured against predetermined criterion levels. If the responses given by an individual student indicated that a particular objective had not been mastered, the teacher would then be responsible for the development of a remediative domain grouping.

> *I strongly suggest that computer technology be employed in the construction of criterion-referenced tests.*

Norm-referenced tests are geared to produce variant scores in which half of the students in the norming group give correct responses to an item while the other half answer the items incorrectly. The items thus chosen are highly correlated with native intellectual ability and do not detect the effects of good pedagogy.

Traditionally, norm-referenced achievement tests have been given at the end of the school year. Criterion-referenced examinations should require pre- and post-testing schedules to help define the students' competencies and detect the effects of good instruction. The post-testing quality could be used as an assessment of scholastic growth.

Great care must be taken to assure that criterion-referenced tests are not affected by the degree of cultural unfairness and bias that has often been ascribed to norm-referenced measures of reading achievement. The criteria that should be selected must not point out differences between subculture groups and middle class society. Rather, the language of the instruments and their content should only represent the influences that are being tested.

The creation of a well specified set of tasks which would be included in the domain groups would not permit the employment of the empirical item selection procedures that have been used in the past. The influence that the quality of instruction has on the student would determine the success that the selection had in measuring the achievement of the specified objective.

The absolute information revealed by criterion-referenced testing would be independent of measurement practices as they currently exist. Properly documented field research is needed to determine the reliability and validity of such devices. Research should commence to define new techniques of empirical analysis that could be used.

I strongly suggest that computer technology be employed in the construction of criterion-referenced tests. The ability of computers to classify and store an infinite number of test items, written at numerous levels of difficulty, would permit educators to develop comprehensive sets of objectives. Statistical devices used for item sampling could then be applied to produce more representative tests from the available objectives.

If they were consulted, most parents of inner-city children would probably say that they are not concerned with the content of reading achievement tests or the ways in which they are graded. Rather, they would wish their children to be spared from the traumatic results of unfair labeling derived from test scores. Most educators would readily agree that standardized testing should not be the sole indicator of a pupil's scholastic attainment. Growth can be demonstrated through children's adjustment to school life, their participation in class activities, and their increasing ability to absorb and comprehend the materials presented to them.

Achievement tests have a valid purpose in program evaluation. But, unless they can be used as instructional tools which enable the classroom teachers to develop groups and to provide pupils with remediation, they will continue to be mere political and social devices for ranking and classifying groups of students and their schools.

Note

1 A. N. Hieronymus, "Evaluation and Reading: Perspective '72," *The Reading Teacher* (December 1972), p. 264.

4. *Teachers on Testing*

Myrna Cooper
Director, New York City Teachers Consortium
and Maurice Leiter
United Federation of Teachers, New York City

We will attempt within the narrow confines of one chapter to confront a significant range of issues concerned with testing. We support the view that the primary function of tests is to contribute to the advancement of learning. We will look with a critical eye at current testing practices but will seek to examine the positive potential of tests and to suggest directions for research, for demystifying the test process, and for increasing the relevance of testing to classroom experience. We will emphasize the importance of teacher involvement in test development, research, and application to instruction. We hope that this endeavor will accomplish at least two goals: 1) reinforce the use of constructive tests in relation to the total educational effort and 2) maintain a perspective about the use of tests so that they be regarded as neither panacea nor scourge.

What teachers want from tests is something other than headlines or stigmatization or celebrity. They have certain convictions about what tests should do. Tests should serve to verify the extent to which what has been taught in the classroom has been learned. Without verification, teachers cannot proceed in communicating curricular content with any rational expectation of completeness or success. Tests should also incorporate learning goals for a given area of study. This obliges the setting of standards requiring clear definition of what is to be learned. Because standards must be grounded in the actuality of the classroom and because they can only be expressed through curriculum development and implementation, such an approach to the formulation of learning goals requires decisions as to how the content of instruction is to be transmitted. Taken in tandem with other assessment clues available to teachers, a test which contributes information about the student being tested can assist in planning the direction and content of instruction on an individualized basis. Thus, good tests as teachers conceive of them can verify learning, assimilate standards, and facilitate adroit decisions affecting the course of instruction.

In the minds of teachers, these instruction-related functions are appropriate criteria for the evaluation of tests. One or more of them should be within the capability of any given test or its rationale will be in question. Within the context of these general criteria for useful tests, teachers,

---

Various illustrative details and a section on minimum competency testing have been deleted from the full version of this chapter –*Ed.*

drawing on concrete classroom experience, have practical expectations which they want tests to fulfill:

- Teachers want tests to provide discrete information about individual performance. They want basic *product* information: what a child knows and does not know. In addition, they need *process* information: why the child hasn't grasped a particular concept; where the conceptual process broke down.
- Teachers want standardized tests to yield more information at the upper and lower limits of the measurement range, for, to be more effective, they require insight into the sum and quality of both excellence and deficiency.
- Teachers want tests to reflect what is actually taught rather than what is written in a possibly obsolete (or archaic) curriculum conceived by those furthest removed from the classroom.
- Teachers want tests to be administered with due consideration of timeliness, and they expect feedback of test information to be available when it is appropriate for instructional needs; otherwise, potentially valuable information expensively obtained is wasted. The fact that tests are given at the wrong times and that test feedback is received at the wrong times is evidence that the instructional function of tests is not a priority.
- Teachers and students would benefit from a different reporting of test results to include detailed item analysis and to yield information capable of illuminating the nature of the errors. Such information would enable teachers to determine whether the errors were group or individual misperceptions and, further, whether they resulted from faulty instruction (material not adequately covered) or from the absence of instruction (material tested which was not part of the course content). Judgments would be possible as to whether the errors were of product or process.
- Teachers feel that tests should be suitable to the age group being tested. For example, a second grade reading test should not involve 78 minutes of continuous concentration. Power tests are, in general, preferable to speed tests for learning purposes. Rate of response is not a measure of learning.
- Teachers administer tests but are rarely permitted to become familiar with the test instrument itself—its purpose, format, and rationale. If they were, they could effectively anticipate aspects of difficulty inherent in the administration process for a particular test. Test security must not be given greater importance than intelligent test administration. Professionals should not be denied access to tests as if they were potential felons.
- Teachers want tests in which design does not mediate the perception of content. If a pupil is being diverted by complexity of directions or of internal format or any aspect of that document other than its substance, the test has failed its purpose. Test format or question presentation should not be an aspect of assessment or a factor in the measurement being sought.
- Finally, teachers want tests normed on the basis of a sample which produces a content validity akin to the real world classroom's curriculum. Often, the norming of standardized tests precedes establishing the standard to be normed. Moreover, these norms constitute a statistical puree the flavor of which changes as the social-economic-educational-political condition of society changes. Like Webster's Third International Dictionary, which tells you not what a word *should* mean but only what it *happens* to mean at the point in the life of the lexicographer as the volume is readied for the press, norms on standardized tests function as passive descriptors on a ship of state without a standard floating from its mast. The test should reflect educators' perceptions of what should be learned and *is* actually

being learned at a given developmental stage. Materials prepared by testmakers should reveal standards underlying the norms to permit test selection appropriate to the learning expectations of the school system.[1] Thus the use of tests would yield something more than *neat* equations. There would be a relationship between the test, those achievements for which measurement is sought, and actual classroom activities. Norms should inform, not obfuscate.

As teachers desire having test instruments perform as integrated components of the instructional process, so, too, they are concerned about the frequency with which tests and test scores are put to uses which are irrelevant to instruction and to its improvement or actually damaging to students, to public confidence in education, and to teachers themselves.

The misapplication and misinterpretation of test results can injure individual students and erode curriculum and instruction. Test scores thus used create social and intellectual segregation, foster elitism, fashion a punishment/reward syndrome, reduce learning to rote and regurgitative modes, deprecate, stigmatize, exclude. To a large extent, school personnel and institutions, reacting to outside pressure and needing test scores as a crutch, have made such practices a part of the fabric of every child's education.

The most serious problem ...
has been the miseducation of the public.

Perhaps, ultimately, the most serious problem resulting from the exploitation and abuse of crude test data has been the miseducation of the public. Critics of schools cite published test scores to argue that we know that schools have failed because tests have told us so. The "telling" consists of a presentation of gross test results and gross interpretations presented through the mass media without refined, sophisticated, or knowledgeable guidance in comprehending the data. Such use creates a public perception of test information in the form of a deficiency model; i.e., testing exists to place blame. The preoccupation of the public rests mainly on two indicators: reports of normed performance of schools and school districts related to "national averages," and individual reports for a student which elate or alarm that child's parents.

The rationale for this type of reporting is given as the public's right to know whether the schools are doing a good job. What the public does not know is the extent and dimension of uncertainty attached to the measure, its meaning, and the environment of variables contributing to the result. What the public does not notice is that the score does not give any insight into the specific learning accomplishments or gaps in learning of a particular student, or into the process of ratiocination which yielded right or wrong answers, or into the nature of difficulty that the student is experiencing with certain classes of items. The score, in short, is merely a gross comparison of one student with a class of students in the same grade and tells little or nothing about any individual that can be used to affect that individual's instruction. Yet it is precisely this kind of information that has precipitated the current sense of crisis in education. Each time a newspaper reports quantified test results, alarms sound for that proportion of students whose ranking falls short of the norm. Deficiency makes for headlines and headlines make for distortion.

Consider, in contrast, the following statement[2] from the National Assessment of Educational Progress reporting on the comparative testing of 9-, 13-, and 17-year-olds in 1971 and 1975:

> During this period we find that the reading ability of American students has changed, but this change is neither all positive or all negative. The results released in this report disagree with the image created by recent publicity surrounding the declining SAT and ACT scores. As I'm sure you know, both SAT and ACT are designed only to predict the success of college freshmen and thus discriminate among and between the students taking them. They were not designed to measure a student's educational progress [or] ....the nation's educational progress.... Recent reports on various test score declines have spurred the American public into discussion and debate about education in this country. If our education system has specific problems, we need to know their scope. We need to know what approaches work...which techniques fail... whether we really should go 'back-to-basics' ... [and] those making the decisions should review *all* available data with a complete understanding of the data's intended use and the data's limitations.... [Thus] we may avoid costly errors in the allocation of education resources....The reading data reported...show significant gains for 9-year-olds while the reading ability of 13 and 17-year-olds remains about the same.... Using National Assessment information and complementary data—the decision may be to continue the current primary grade emphasis while expanding the reading programs to serve junior and senior high students.

> *... the possibilities for tests are greater than the obstacles to their realization.*

The point is that the controversy relating test scores to school performance exists without a sound basis in fact and is largely a product of superficial or misinterpreted information. And where data are offered which suggest a "real" decline in achievement test scores, they are usually accompanied by vagueness as to the cause of the ostensible decline: "We will not be in the position to prove any causal relations, but we will show the potential importance and power of some factors."[3] This is why those who are sensitive to the use of tests as the basis for attacks on schools place heavy emphasis on the relative imprecision of standardized measures, on not concluding from their data what their data have not been developed to conclude, and on the importance of factors not related to school performance (such as population mobility and poverty). None of this conflict is helpful to teachers, for we are concerned with what tests tell us about the specific children we teach. In the process of dealing with attacks on schools fueled by the public obsession with standardized scores, school districts and teachers are often obliged, out of concern for the survival of students, their own professional reputations, and public education, to place undue emphasis on test success and to treat test performance as an end of education or an end in itself.

It is most unfortunate that the testing movement has gone away from an emphasis on assessment as a part of the instructional mode and that we, as educators, have been overcome by a misguided politicization of the test mechanism. This is particularly regrettable because the possibilities for tests are greater than the obstacles to their realization.

However, we do not believe that tests can be made more relevant to instruction in a vacuum. To be part of any instructional process implies not only integration with curriculum development, strategies of teaching, and evaluation, but a relationship to a framework within which acted out notions about the purpose of schooling and the potential of learners are acted out. What and how we teach and why and how we test reflect our convictions and our commitment. Perhaps the way to make testing more relevant and effective is to arrive at an understanding of the meaning of what we are doing.

Consider, for example, the instructional approach popularly known as Mastery Learning. In his recent book, Benjamin Bloom states, "Societies in the past have relied largely on prediction and selection of talent as a means for securing a small group of educated elite.[4] Modern societies," he concludes, "stress the development of a very large number of well-educated persons and attempt to produce this by legal and social pressures" (such as compulsory and compensatory education). Since our society places such great value on education, it must develop educational strategies which will make school more meaningful to individual pupils and must find the means to develop talent.

Moreover, Bloom is critical of the current use of standardized achievement tests which make judgements on whether a pupil has learned or will learn more or better or less than his/her peers. Bloom claims that scholars, researchers, and others, expecting differences, design assessment instruments which will provide theoretical, experimental, and practical justification for the process in which they believe.

Bloom refutes the popularly held view of achievement as a function of aptitude which is fixed, stable, and observable and which, at each stage of schooling, typically shows greater individual differences in learning attainment than was true at previous stages. He is critical of those who would use the concept of individual differences as measured by achievement tests as rationalizations for depriving some students of the opportunity for further learning. He has drawn attention to a growing body of knowledge which indicates that almost all children can learn what the school has to teach if provided with proper learning conditions.

Because teachers have been trained to believe in achievement as an entity which falls rightly onto the normal curve, they consider a unit, regardless of its origin (syllabus or published works) to have been successfully taught and learned if the test results at the conclusion of the unit fall into the established pattern of the normal curve. Pleased with the curve of the normal distribution, because statisticians tell them that it's supposed to be this way, teachers often do not provide correctives to those who naturally fall below the grade norm. Consequently, approximately half the class proceeds to the next learning not quite prepared to acquire new skills. Bloom's approach to learning seems to call for a reexamination of curriculum, instruction, and testing as we have traditionally known them.

The rationale and approach of Mastery Learning illustrate a number of important points about testing. For one thing, Bloom's view is that using a structured and sequential learning method will result in students achieving test results wherein the regular distribution of scores will be bunched at the top of the scale. This is in line with what teachers really want to accomplish: namely, the success of all their pupils. In addition, this approach underscores the importance of the teacher-made test as an essential tool of assessment, particularly in the short range. For a variety of practical reasons and obstacles, standardized tests are not now usable for obtaining information on pupil performance for other than long-range periods and broad areas of knowledge. The gap is filled by teacher-made tests, but the criteria are really no different. One still requires items which reflect

what is actually taught, which, on analysis, will reveal what has not been learned, and which must be ordered and sequenced in relationship to the curricular goals. Most important, these tests are used as part of the continuum of instruction. In fact, the Mastery Learning strategy is inconceivable *without* testing as an internal component. Therefore, while we continue to require long-range assessment tools to evaluate relative adherence to a standard and to guide in decisionmaking, we begin to see that, within the classroom, the teacher-made mastery test is perhaps the most important assessment instrument.

Research and Development on Testing

There is a great deal to find out about learners and learning, much of which can be derived from worthwhile research and development on tests. The methodological and technical problems associated with test construction and feedback are under control, and we are, in a sense, at a juncture in measurement at which it is reasonable to assert that tests can be made to do almost anything we want them to do. We lack not information but direction for testing and integration of the testing information currently and prospectively available. Unfortunately, research in the field tends to serve primarily the research community itself; it tends to be introverted and to be largely incomprehensible and inaccessible to others in the field of education. This is not solely because the research is recorded in an abstruse metalanguage. A piece of research worth doing can be described and explained in the common tongue and can be made meaningful to intelligent and interested educational personnel. However, at present, the wall between researcher and practitioner does not promote teacher confidence in tests or in educational research as a means to improve teacher effectiveness.

> *We lack not information but direction for testing.*

A similar problem exists in the specific area of test research and development. There is a gulf between testmaker and testgiver, between test and curriculum, between test and actual classroom activities. There is a substantial lack of articulation between practitioners and researchers. Teacher input into the content and purpose of tests is rarely sought. Idealized assumptions are made about curriculum and about whether the knowledge being tested was conveyed and whether the learning was *in* the classroom.

The research and development aspect of testing would be enhanced by articulation with practicing teachers in the course of test conception and construction. The use of "experts" is insufficient since these people are frequently removed in time and place from classrooms. In many instances, the curricular basis of test items is twice removed from reality, taking its cue from the testmakers' published curriculum materials which are themselves idealized and probably generalized versions of the actual classroom's curriculum.

If teachers are involved in test development and use, and if research results are communicated to them, teachers will be encouraged to acquire greater sophistication in the use of test instruments and to place greater faith in test information and general research as applicable to classroom instruction. While "the daily challenge of the classroom simply does not demand...psychometric elegance" it is clearly not an impetus to the advancement of our thesis that measurement is an integral part of

instruction to have between three- and four-fifths of practicing teachers estranged and virtually helpless where measurement is concerned.[5]

Colleges of education bear part of the responsibility for this situation and for its correction. In teacher education, the tendency has been to present measurement as essentially a statistical subject with some attention given to familiarizing teachers with specific versions of commonly used affective and achievement tests. To make measurements meaningful, however, teachers need to know how to construct tests in an efficient manner using the materials available to them and meeting the developmental needs of specific pupil groups. They need training in developing test items to yield product and process information, a notion of how to interpret test results, and an awareness of informal assessment means such as oral questions, logging, and observation of pupil behavior. In addition, testing must be presented in connection with training in curriculum development, instructional methods, and child development. The pieces of the educational complex should be, but are rarely, integrated into a whole.

In considering useful directions for research and development in testing, we are conscious of the importance of making research priorities. If it is our conviction that most of what schools have to teach can be accessible to the vast majority of students, then research in testing should reflect that belief.

If it is our conviction that the primary role of testing is to inform teachers and contribute to instructional decisions, then research and development should be directed to making tests better perform that function. If it is our conviction that research on teaching should not be disconnected from research on testing because testing is a part of teaching, then research should be constructed to study the test as an instructional tool. If it is our conviction that tests need to be more than statistically "nice," and that they should perform a formative function as well, then test R&D should be directed toward refining our capacity to develop tests which reflect specific standards for specific curricula and for real-life classrooms.

We suggest, within these guidelines, that testing research priorities be directed to the following kinds of activities:

- The development of normed power tests to exclude time as an achievement factor (what some pupils need is more time to learn and to reflect on what they have learned—for time may be the only variable in learning which schools can control).
- The development of diagnostic standardized tests which yield discrete feedback capable of prescriptive application.
- The development of achievement tests which yield both a measure of what has been learned and information to provide direction for further instruction to individuals and groups.
- The development of tests which are responsive to our national commitment to meet children's learning needs "where they are at." If we want to do this, then we ought to face the real problems and devise instruments to provide precise information on language assessment, early indentification of gifts and talents, learning disabilities, and physical development.
- The development of tests to measure that whole other dimension of schooling, the affective domain. We recognize that the schools have traditionally had the role of promoting affective learning of children (learning how to learn, human interaction, social participation, formation and growth of self-concept, problem assessments, exercising options, valuing, and the relationship of these to cognition and instruction). Certainly, these diversions are more

pertinent life skills than handling 1040 forms. They form a seminal part of learning which seems to have largely escaped the attention of both educational researchers and basic skills zealots. Beyond some work on cognitive styles, little serious research exists.

- The development of research into testing as a tool in arranging a child's learning environment—testing as a measure of the appropriateness of environments and to verify theories of instruction and cognitive style.
- The development of research into the use of tests as a means to identify schools which are successful with various types of learning problems—research and test development to isolate the components which create the climate of success.[6, 7]
- The development of research to determine why changes in general pupil performance have occurred and whether they are good or bad; e.g., if current standardized achievement tests suggest that 8-year-old pupils today are less efficient at decoding than their counterparts 10 years ago, can we by analysis determine what has yielded this change (departure from phonic instruction, emphasis on context) and, as a corollary, whether there has not been a concurrent gain in an accompanying area such as comprehension? We have too readily assumed that a decline in one area measured signals an absolute rather than a relative learning loss.
- The development of research on reporting of test information to teachers, to parents, and to the public in order to find effective ways of explaining test results, making test functions understandable to a variety of audiences, and making test information reporting comparable in proportion and intensity to other kinds of school performance communications.
- The development of an R&D complex comparable in scope to existing laboratories and centers, with the significant difference being the integration of this one with an existing large urban school system like New York's. Such a center could draw upon the data-base provided by the diverse pupil population, the expertise of the many researchers associated with the large number of urban universities, and the field-based experience of thousands of practicing teachers. It would serve as a living laboratory and as an arena for collaboration and interaction.

Regrettably, a formal place for teacher input into the R&D process is lacking. Teachers' participation and expertise should be reflected in every stage of research, from conception and design to execution and evaluation to dissemination and application.

In connection with these specific recommendations on research and development, we have some views on the burden of testing and the responsibility to disseminate information that research has yielded which are related to the role of Government in this area.

Clearly, the financial impact of testing as a function of evaluation or qualification for Government funding is more than school districts can bear. It is, ironically, a regressive tax on those districts least able to pay. Federal or State Governments should assume the cost for tests which districts are obliged to administer to receive or retain funding. In any event, it is certainly possible to find less expensive ways to obtain data not related to individual instruction (estimates of general achievement, qualification for funding, etc.) and to simultaneously reduce the physical burden of testing. The use of matrix sampling is an obvious approach.

Finally, we are troubled by the absence of teacher recognition of institutional forces in the field of research and by the lack of articulation with teachers by those who "create" the literature of research. Few teachers are aware of the existence of the National Institute of Education; or, if they

do know its initials, they know little of its responsibility and scope. Nor does the existence of ERIC offer significant benefit to teachers, as it is mainly useful to researchers, scholars, graduate students, and proposal writers. We suggest that NIE consider passing up two or three of its more recondite grants each year and use the money thus realized to disseminate directly to every practitioner in the country a concise publication titled, perhaps, Research of Use to Teachers. By this means, and by standards for research and interaction with the world of the teacher which NIE is in a position to set through its granting authority, significant progress could be made in promoting teacher participation in developing and applying research on testing. In addition, NIE should consider setting aside money on a minigrant basis to fund teachers directly in research projects of their own.

Few teachers are aware of the existence of the American Educational Research Association. Despite the fact that teachers are looked at, measured, investigated, observed, and frequently judged by the numerous thousands of members of the research community and despite the fact that the purpose of educational research is application for the improvement of teaching and learning, there is no formal interaction between researcher and teacher. Were any teachers among the 10,000 in attendance at the last A.E.R.A. convention? We feel it is important to formalize the relationship, to create a means of obliging researchers to reveal their work "in the common tongue" to teachers and of encouraging teachers to react to plans for research, the results of research, and the applicability of research. We suggest a series of national and regional interaction conferences to promote relevant research and meaningful dissemination through close encounters of the two constituencies.

While we believe it to be valuable—in fact, essential—to make the practicing teacher an organic partner in the activities of test research, construction, utilization, and analysis—a partner in this as in any other aspect of instruction—and while we consider it helpful to seek the personal and concrete views of teachers on this area of assessment, we do not believe that teachers can overcome the intrinsic and extrinsic difficulties and shortcomings which tests and their interpretation currently present in isolation as individuals or even as a single body within the profession. The solution to our problems with testing requires coordination among many groups in and out of education, including but not limited to higher education personnel, researchers, political leaders, the public, and the media.

Notes

1. Standardized tests tend to be too closely linked to the curriculum publications of the companies which sell both tests and teaching materials. To that extent, commercial bias can influence both curriculums and tests.
2. Roy H. Forbes, *Statement of Change in Reading Achievement of Young Americans*: News from National Assessment of Educational Progress (Denver: Education Commission of the States, 1976).
3. Annegret Harnischfeger and David E. Wiley, "Achievement Test Scores Drop. So What?", *Educational Researcher* (March 1976): 5-12.
4. Benjamin Bloom, *Human Characteristics and School Learning* (New York: McGraw-Hill, 1976).
5. Benjamin Rosner, "Teachers' Perceptions of their Tests and Measurement Needs," in Karl Heinz Ingenkamp, ed., *Developments in Educational Testing* (New York: Gordon & Breach Science Publishers, 1967).
6. Charles I. Schonhaut, *Accountability Program Progress Report* (New York: Board of Education, 1978).
7. Garlie A. Forehand and David E. Wiley, *Implementation, Usefulness and Consequences of Minimum Performance Standards* (Princeton, N.J.: Educational Testing Service, 1978).

Principals

5. *Standardized Testing in Elementary School: A Practitioner's Perspective on Several Significant Testing and Evaluation Issues*

Parker Damon
Principal, McCarthy-Towne School, Acton, Massachusetts

This chapter focuses on: (1) the retention and distribution of test information, and (2) the interpretation of test validity. Both are issues an elementary school principal confronts and has a professional responsibility to handle in a way that protects and serves the best personal interests of individual students and their families, the professional rights of faculty, and the pedagogical concerns of the school and district. Unfortunately, the use of standardized tests makes it difficult if not impossible for a principal to fulfill this responsibility adequately and equitably to each constituency.

I have seen the results of standardized tests[1] used to make numerous administrative decisions:

1. To transfer a principal.
2. To compare the effectiveness of the total curriculum of different schools.
3. To determine the need for new materials.
4. To gage the relationship of curriculum goals to particular materials and methods.
5. To evaluate teacher effectiveness.
6. To diagnose an individual student's strengths and weaknesses in particular curriculum areas.
7. To determine admission to private school.
8. To place students into ability groups.
9. To predict an individual student's performance.
10. To reinforce budget priorities.
11. To compare individuals and groups of students to each other.

Frequently, the same test battery supports many, if not all, of these decisions.

Sections stating the author's views on the influence of tests, a professional's responsibility regarding test content, and various illustrative details have been deleted from the full version of this chapter–*Ed.*

I know it is not possible for any one test to do all these things in a way that makes sense. Therefore, I believe a principal has the responsibility to try to prevent this kind of misuse. I have found, though, that this is practically an impossible task.

The back-to-basics and competency testing movements illustrate and intensify the increasing general interest in the use of tests, regardless of their faults.

Despite what testmakers state about the proper use of their tests, professionals and laymen alike place unwarranted value on test scores. People in both groups misinterpret and/or misapply test data even though they have been warned by professional and citizens' organizations. The traditions of elementary school emphasis on the student and secondary school emphasis on subject matter encourage some of this misunderstanding. The back-to-basics and competency testing movements illustrate and intensify the increasing general interest in the use of tests, regardless of their faults. Teacher education institutions do not stress the critical analysis of test content or comprehensive "determination" of the appropriate application of results. Some of us who believe there are alternatives to using standardized tests find it difficult to be heard; we are rarely taken seriously. Proponents of the proper use of tests need help. Parents, teachers, administrators, and school board members cannot cause the proper use of test results to happen all by themselves. Even when they work collectively, their efforts are eventually overpowered by the inertia and size of the total educational complex. The experiences my colleagues and I have had over the past few years at our school portray some of this problem.

A Case History

The McCarthy-Towne school has an enrollment of about 450 students in grades k-6. It is one of five public elementary schools of equal size serving an upwardly mobile, middle class suburban community 25 miles west of Boston. The school was started 8 years ago as an alternative "exploratory" school. At that time, the school enjoyed the strong backing of parents, some school board members, and the superintendent. It was also the target of a lot of criticism. Among aspects of the school that were attacked were its methods and materials for the teaching of reading and math.[2] As one means of dealing with the conflict, the school board decided to use criterion-referenced tests to determine whether the school was doing "at least as well as the others."

As a public alternative elementary school, McCarthy-Towne differed at the start from the other four schools in several ways. Faculty members knew when they were hired that they would be part of all decisions; each, like the principal, had one vote, and only the superintendent and the school board had the usual veto power. Faculty and parents agreed that a quality education could, for them, be best attained without recourse to certain specialized staff positions (art, learning disabilities, music, physical education, and remedial reading teachers) or reliance on texts and workbooks. As a result, each classroom teacher had the responsibility for all areas of the curriculum. This was made possible by reallocating much of the money not used for some of the specialized staff for other purposes, such as a coordinator of volunteers and student teachers, consultants, and inservice

workshops. Teachers had a lot of help available if they wanted it, but not at any added expense to the school's budget. Through the inservice sessions, the faculty became adept at using the Gattegno approaches to teaching reading (Words In Color) and math (Algebricks) as well as in developing a different approach to teaching all the other subjects. Teachers knew that their control over the school's curriculum and learning environment was complete, but not supreme. They sought and received parental and district support. Now this educational entrepreneurship is being eroded by the pressures of competency testing and the like.

These tests were created by representatives from each school with the assistance of consultants from a nearby college. Although the results were inconclusive, a local newspaper used certain subtest scores to show that indeed our school was not only doing "as well as" but "better than" the others. This misuse, plus faculty and community feeling that the tests were either too easy or did not accurately reflect the curriculum, soon brought about their being discontinued.

The superintendent formed a committee of teachers and administrators to establish the criteria for measuring achievement and the tools for doing the job. The committee did so by surveying what was taught in each school, trying to match or create test items that reflected this teaching and each school's materials, and polling each faculty member for opinions about appropriate levels of difficulty and standards to use for establishing successful achievement. There was not enough time to do all of this well, a fact that nearly everyone realized once the project was begun. The tests which were developed at the end of a year's work omitted significant parts of the curriculum being evaluated and distorted some other parts (if students work with numerical fractions but not fractions of objects such as shaded shapes, can the one form be used to evaluate an understanding of the other?). Much of it was still too easy (when there was disagreement there was a regression to the use of easier test items). These particular criterion-referenced tests did not challenge the students to show what they could do, nor did they accurately determine where students were weak, or whether or not any particular programs were more effective than others. Another problem, one not publicized, was the fear by some faculty members throughout the system that they would be evaluated on the basis of how well their students scored.

Because of the above reasons and because they desired to be able to compare the performance and potential of students in one district school to another as well as the students within the district to those outside, the district adopted the California Test of Basic Skills (CTBS) and the Short Form Test of Academic Aptitude (SFTAA) battery for use with all grades except kindergarten. They are now given every other year starting with the second grade.

The data from these norm-referenced tests are as worthless as those from the criterion tests, but for different reasons. The content of the criterion tests could have been redesigned to be more valid. However, the CTBS test purports to measure reading and other language arts abilities, for example, that an analysis of individual test items shows it does not measure.

Because we at our school have felt so strongly about the inadequacies of the standardized tests, we have been challenged to offer reasonable alternatives, and we have done so. However, we have been unable to bring about any significant change in testing practices.

The situation in Massachusetts is further complicated by two State laws which influence the use of test results. The Massachusetts State Board of Education is moving toward the adoption of a statewide minimum competency testing program, so there will soon be a third law. (Since this was written, the State of Massachusetts has passed such a law–*Ed.*) One of these laws regulates the way in which students' records are kept and used. The other specifies the procedures to be used to

identify students with special needs and the procedures for providing any special programs.[3] Both are good laws intended to protect students' individual rights and to equalize their educational opportunities. Both, however, refer directly to standardized tests. As a result, the references to standardized tests by State and Federal regulations are a mixed blessing. On the one hand they seek to protect students, while on the other they tend to support, by the mere reference to the tests, an educational procedure which may cause abuse.

Along with proposing alternatives, we have tried several other tactics to reduce the school community's reliance on test scores. For example, each year we write a report of what the latest results mean and distribute it to faculty, parents, and school officials.

Second, we have corresponded with the publisher of the tests we use in order to learn how test items are actually constructed. Our intent was to discuss with parents the tests' poor content validity. However, the publisher did not cooperate with this request.

A third tactic has been our school's participation with Project TORQUE,[4] which is developing a new kind of test. We have been able to show that the CTBS tests misidentify high and low achievers in math computation, and that areas of achievement labelled satisfactory really are not. The point we tried to make was that if one part of the CTBS was providing incorrect information, why should anyone put faith in the data other subtests supply?

A fourth tactic has been to withhold test results from students' records when they are passed along to the junior high school. We have used different methods to do this, and have been only partly successful.

Some people have become more "test conscious" and they are questioning the tests' content validity, diagnostic value, predictive capability, and summational usefulness. However, these reservations are diminishing under the pressure of the growing popularity of competency testing, special needs assessments, identification of the academically gifted and talented and, in our school, requests for report cards and grades.

What Is the Professional's Responsibility To Retain, Distribute, and Interpret Test Information?

As mentioned earlier, Federal and State laws govern much of what school people may or must do. But within this framework there may be a lot of leeway for principals to exercise leadership. For example, in our State the student record consists of a transcript and a temporary record.[5] The latter may include standardized test results and may be periodically reviewed by a principal for the purpose of destroying misleading, outdated, or irrelevant information. The principal has a responsibility to weed out old test data, and should be encouraged to do so from outside the school. In addition, elementary school principals or their designees should meet with parents to review the contents of each child's student record prior to its being passed along to the next level of schooling. Regardless of who carries out this work, the process should involve (if it has not already been done) an examination of the actual tests, the actual answer sheets, and the scoring and interpretation forms. The scoring and interpretation forms only become meaningful in the presence of the other two. Again, support from outside the school and district would encourage this kind of activity and help overcome the attendant obstacles of time in which to do it, space for keeping such complete information, and money for possible additional faculty time.

Along with the responsibilities to parents, pupils, and other professionals to insure the proper use of test data, principals must also acknowledge and fulfill certain obligations, prerogatives, and expectations.

Principals are often obliged by their own contracts, or those of teachers, to hold meetings and perform other duties within set time periods. This contractual obligation, when coupled with legal requirements to conduct and complete educational needs assessments according to strict deadlines, can mean that testing is hurried and not as thorough as it should be. For example, it is not easy to bring together a team of professionals for early morning, late afternoon, evening, weekend, or vacation meetings to discuss assessments and educational plans with parents. The end of the school year and then the summer vacation before school begins are often excellent times for doing this kind of work, but contracts and costs together prevent this time from being used. Therefore, principals may be forced to function in other ways that conflict with the responsibilities that have been mentioned.

Or, the principal may not be allowed to keep test booklets and answer sheets with the student record folders, but may instead be required to return them for safekeeping somewhere else. Or the principal may be prevented from handling pre- and post-test information differently from the way in which the district does. Thus the principal may be prevented from giving additional information or advice.

> *Principals need help in taking advantage of and exercising their prerogatives.*

Because principals occupy a unique position in the educational system, they hold prerogatives not available to other educators. They have access to every classroom and are privy to student and parent concerns, involved with teachers' curriculum problems, and informed about district and community issues. As a result, principals should take stands on what a class, school, or district should do, and they should express opinions about the merits of philosophies, policies, and practices. My experience is that principals are, on the whole, a quiet group who do not speak out or exercise prerogatives such as these. Or if they do, they fail to make any significant large-scale impact for want of appropriate technique and training. Therefore, principals need help in taking advantage of and exercising their prerogatives. And a district's student and program evaluation process is a good focus to have when working on this revitalization.

Principals, like everyone else, assess what they do in light of their own, their family's, and their community's expectations. Are they fair, open, and curious about new developments such as those related to testing? Do they seek out new information, work to help others acquire new information, and share the information they have with others? For example, will a principal recommend for hiring, rehiring, or dismissal someone who does not know very much about testing and its implications and alternatives? Principals might be helped by research showing how their professional judgment is affected by the pressures of these different expectations. The topic of what to do with test information provides a stage on which to examine these adversary and supporting roles.

Standardized tests do not deal with several key factors:

1. Standardized tests get in the way of the teachers', schools', and districts' responsibilities of setting the criteria by which programs and groups of students will be evaluated. It is not just that the norming and comparative data obstruct any predetermination of what the criteria should be. It is also that many professionals and lay people assume that the test-makers have already established what the standards should be and what the best ways are for finding out if these standards have been met. I believe tests have far more credibility with parents than with teachers and principals. Thus, professionals have a hard time getting other standards and criteria accepted by the public. Currently, standardized tests do not help with this problem.
2. There is no list of points on which to evaluate the tests themselves. Such a list, plus a comparison of test and curriculum content, would do much to improve the way in which tests are used.
3. If there were such lists for parents and professionals to use when evaluating the tests themselves, it still might be very difficult to have the tests available to examine. Most testmakers protect from public scrutiny both the details of test design and the content of the tests. Testmakers argue that unless the tests are kept "secure" the norming or comparative value of the tests will be impaired, and that some students will acquire an unfair advantage over others. They also claim that the cost of developing each item is so great as to preclude the production of an unending supply of alternatives.

   The test secrecy issue poses some interesting problems. On one side are the testmakers who wish to protect their product by means of copyright laws and trade secrets acts. On a second side are the Family Educational Rights and Privacy Act (the Buckley Amendment) and discrimination laws that provide students and parents with access to information being accumulated about them.

   Still a third position is represented by the Freedom of Information Acts which permit individuals or groups to examine the records of governmental committees and to observe their deliberations in open session. If a taxpayers' group, for example, wanted to challenge a school board's decision to allocate huge sums for wide-scale testing, their case would be greatly strengthened if they were able to compare and openly critique the tests under consideration.

   Lastly, there is the moral side of the secrecy issue. Should not all the parts of public education be public? Many projects funded by the government require the use of private tests, and thus the government is both subsidizing the testmakers and supporting their reliance on secrecy. Not surprisingly, the testmakers are content to overlook these problems. Thus, they maintain the status quo which, in turn, translates into the continued misuse of tests.
4. Standardized tests do not take into consideration the way in which many professionals and parents view schooling. In elementary school, especially the early years, teachers are student-centered as opposed to subject-centered. Standardized tests do not foster this developmental, holistic, and interrelated view of students and schooling. Instead, standardized tests foster the compartmentalizing of learning that occurs in secondary school, the separating of abilities and interests into hierarchies of importance that starts in the middle grades, and the comparing of individual performance to group norms that begins with kindergarten. The only significant difference in the way a 12th grader and 1st grader take a standardized test is that 1st graders have theirs read to them. Shouldn't different stages of schooling have different types of tests? Shouldn't both content and format differ?

5. Standardized tests confuse quality and quantity. They are not designed to show how well a student can perform a particular task, but only with whether or not a certain number of tasks can be done; they are not concerned with how an idea may be expressed or interpreted differently.

   Student productivity is the heart of elementary schooling. Teachers in elementary schools try to harness children's proclivity for activity by using interrelated experiences, physical involvement, and assorted technology. They hope for productive activity, not busywork. National and state assessments focus more on what students actually do than do standardized tests. In addition, standardized tests direct attention away from more important assessment and evaluation issues.[6] The kinds of questions students ask, how they seek and give answers, and the way in which they choose to describe and explain ideas, emotions, and events should be part of the information acquired by standardized testing. Too much effort is devoted to too little return, and there is too much stress on one correct answer rather than an order of acceptable ones. Perhaps the NIE could focus attention on the development of tests that explore what quality means in the elementary grades and why alternative measures of performance are needed at both elementary and secondary levels.

6. There is a general need for more widespread understanding of how standardized test results correspond to other forms of evaluation. Some work has already been done in this area, but a lot of confusion still exists about whether various types of measurement instruments are actually different. For example, is a teacher-made test any more/less subjective than a commercially made one? How does the content of norm-, criterion-, and domain-referenced tests differ? Questions such as these need more attention if there is to be greater trust between school and community.

A Vision

Standardized tests do provide a lever on the vast educational system. This is a lever that can force, direct, and assist improvements in the learning experiences of all students. Therefore, any restructuring of tests should consider their present and possible uses. For example, I believe that the Federal Government, by intervening with improved evaluation and assessment practices, could cause the following to happen within 10 years:

1. An employment literacy for 85 percent of the student population by the time they graduate. That is, they would be able to read high school materials designed for 10th graders.
2. There would be no reading failures.
3. The dropout rate would decrease significantly since none would occur as a result of not being able to read.
4. The quality of literacy for all students would be improved and observable. Students would not only be doing more reading and writing, but the quality of both would be improved. Library use would expand as would sales of books.
5. There would be no school-caused boredom. Everyone should be allowed to bore themselves some of the time if they wish to, but schools would no longer be boring students with unnecessary repetition, dull materials, phony situations, and low expectations.
6. Students would correctly perceive that experiences in school, or those which are school-related, are at least as exciting-interesting-challenging as those occurring out of school. Student newspapers, radio stations, TV production, moviemaking, theater, dance, and

music performances on all scales, experiments and explorations of all kinds, both at the secondary and elementary levels, would occur as a result of the improved use of tests. Time now spent on unnecessary activity would be available for other forms of activity not now seen in schools.

All this is not an impossible dream since each is already a reality in more than one spot. A vision made up of elements such as these will become reality if the Government exerts its power.

Notes

1. As used here, the label "standardized tests" describes *both* group-administered norm-referenced achievement tests *and* group-administered norm-referenced aptitude or intelligence tests.
2. See Caleb Gattegno, *What We Owe Children, The Subordination of Teaching to Learning* (New York: Avon Books, 1970); and *Towards a Visual Culture, Educating Through Television*, (New York: Avon Books, 1969). These two books give brief descriptions of the pedagogy used by the schools.
3. These two laws are similar to two Federal ones: Education of the Handicapped Act (P.L. 94-142) and the Family Educational Rights and Privacy Act (the Buckley Amendment).
4. Project TORQUE (Tests of Recurable Quantitative Understanding of the Environment) is a foundation-funded project at the Education Development Center in Newton, Massachusetts. The TORQUE approach features specially designed games and activities that yield information about students' mathematical understandings. In each of these games, a child's ability to perform depends directly on his/her competence at a specified set of mathematical skills.
5. *Student Records Regulations*, Massachusetts Department of Education, 1976.
6. Assessment looks at performance to predict what is needed by a student or group, whereas evaluation looks at performance to see if goals have been met and methods are successful.

6. *An Elementary School Principal's View of Standardized Testing*

Luis Mercado
Principal, Public School 75, New York City public schools

Standardized Tests

Standardized test scores are meaningless for diagnostic purposes. There is no way of knowing why a child has selected a wrong answer. Is the error due to not being able to decode written words? Does the child understand the task? Does the child's reference framework make for different interpretations? Are there multiple possibilities in the answers? Is there no best answer? Piaget describes the differences in children's conceptual thinking as they pass from the concrete to the abstract developmental stages. Is this the reason for the error?

It now appears that standardized testing has merged with the trend toward "minimum competency" in teaching reading and basic subjects. In high schools it is reflected as standards to be met by students graduating from high school. In the reading area it becomes laundry lists of skills taught and tested.

> *The effect of the minimum competency movement is to produce an overemphasis on the test score.*

In our district and our school, the effect of the minimum competency movement is to produce an overemphasis on the test score. A thrust for curriculum changes, with an emphasis on isolated skills, is evident. In a recent conversation, one of the highly skilled reading teachers in our school's summer remediation program made a strong statement that the basic problem for our children is lack of comprehension. Her solution is more systematic instruction using basal readers and reading-out-loud activities. I observed her teaching. My interpretation is that the teacher is establishing relationships, promoting language art activities in a small group setting (six children), and increasing

---

Sections on bilingual education, testing alternatives, and perceived testing needs of minorities and the handicapped have been deleted from the full version of this chapter–*Ed.*

motivation as well as comprehension. Thomas Murphy of Holt Rhinehart and Winston said, "Teach the skill—that's the new old time religion. But this makes me uncomfortable. Reading is more than skill acquisition. It is an art form. You can cripple children by not giving them enough literacy."

Publishers are designing and have on the market, built into new texts, tests resembling standardized tests. Is the main purpose of reading to get the right answers on multiple-choice tests? Standardized tests are not useful for diagnosis. Coaching and rote practice activities could very well become parents', teachers' and children's main preoccupation. Do poor test results mean that a child who is reading cannot read? Are some of our reading failures due to the testing process? As we explore the testing area we observe children making errors because they have never learned certain information. Children are penalized because of a lack of knowledge. The test developers' assumptions of common cultural knowledge creates culturally biased tests.

I would like to see the social inquiry case study method used at our school to assess children's literacy. Robert Stake indicates that the case study approach is in harmony with aims of understanding, extension of experience, and conviction of what is known.[1]

I would give children a choice of reading materials. I would interview them using a case study approach. We would be able to determine their comprehension of printed material and their oral reading ability. The case study approach is holistic and difficult. The facts are drawn from interrelated variables. The information is gathered by personal observation and reported in a literature-narrative form. The main focus is understanding the individual case.

I would like to have people who are involved in testing write the tests. Teachers, local test specialists, principals, children, and parents could participate in test construction. We would decide what material is to be learned, and how and why it is important. What are our goals? What is it that will be useful to us after finishing the course? In any event I would not use multiple choice questions because we should test understanding and comprehension.

I believe that our tests could then be related to actual learning goals. Teachers would be trained in the processes of making tests and evaluating learning activities. To involve students in composing questions would enable them to better understand what they have learned. This type of criterion-referenced and meaningful testing is what I have been experimenting with at P.S. 75 (Manhattan), as site director of Broad Jump, a tutorial remedial program based on self-image and motivational activities incorporating the arts in reading and math instructional processes.

Barbara Di Novo, visual arts teacher in Broad Jump's summer program at P.S. 75, shared with me her thoughts on the only test she remembered from her schooling. This involved the type of active learning that I have in mind. "The only exam I remember in 12 years of post-high school education was the final exam during my sophomore year on 17th-century poetry. The professor instructed us to write the preface and table of contents for a book on five 17th-century poets, selecting the best or most representative poets and their five best or most representative poems. This is the only test I remember that asked me to discriminate, to evaluate, and to compare based on a personally developed criterion."

What Should Be Done?

Acquiring skills is not usually a sufficient goal in cognitively oriented learning experiences. I would like to see more mastery experiences, defined as the utilization of those acquired skills. Barbara Di Novo's experience on that one test reflects the premise that there are basic mental

activities underlying her performance, which can be taught and then tested. These are mastery experiences utilizing skills acquired by students and are a credit to the criterion-referenced-type teacher-made test.

Criterion-referenced or objective testing requires full involvement of teachers in selecting and developing objectives. Goals have to be agreed on. What goals are not being accomplished after we've tested children? How do we get students to succeed in achieving missed objectives? I prefer to have our school begin to consider this problem not as remediation, but as a rethinking of high-order reasoning and comprehension. How do we systematically think? Research describes approaches for teachers that develop problem-solving skills. These approaches are based on the premise that there are basic mental activities which underlie performance and which can be taught.

The system used is thinking out loud as both teacher and student work through ideas, analyze relationships, separate concepts, and generalize. If I applied this thinking to my summer activities of directing a remediation program for 90 youngsters at P.S. 75, then our followup school year program for this population could involve the development of criterion-referenced tests. I would involve the teachers, parents, students, and administrators in a dialogue establishing instructional objectives. I'd need technical assistance from evaluation specialists on test item development. I'd ensure that our objectives are consistent with district and State goals for education. I'd think through some possible educational outcomes based on long-range goals for each student. These would require a broad range of test formats. I would not use multiple-choice or true/false questions.

But in any event I'd use a variety of criteria for evaluating our students' learning development. I do realize that care must be taken in selecting objectives. The complexity of developing objectives needs the development of leadership team approaches and is difficult. The payoff for everyone is the improved diagnosis of individual learning problems and the prescription of systematic approaches to strengthen the learning situation.

I am not advocating the development of minimum standards or a pass-fail system. I want our students to be judged on their individual performances. I want our criterion-referenced tests to reflect instructional objectives. In developing criterion-referenced tests, I'd like to correct for cultural and socioeconomic bias. I'd like to emphasize that student tests should not be used for teacher evaluation. This destroys our trust relationships. It demonstrates a lack of understanding of what tests do measure. Tests of any type reflect only a small part of student learning or teacher effectiveness in the classroom.

I believe that criterion-referenced tests where the student is evaluated as to the achievement of goals avoids some of the problems of standardized norm testing. But there are other possible problems. The refinement of structured materials in programed materials leads to frequent testing. The danger is that shorter and shorter test intervals push out teaching. We can end up right where we started, teaching to the tests. Another serious problem in criterion-referenced tests is the local districts' determination of what passing level is acceptable. Is it 30 percent, 45 percent, 60 percent, or 90 percent for students on a task? In any event, the teachers' job is to help the students become more proficient at the task. The criterion-referenced test advantage is that we can relate it much more closely to what is taught in the classroom.

The break-it-down-into-its-parts approach to reading is exemplified by standardized testing. The devisers of these tests believe that a series of discrete subskills or behaviors, that can be independently observed and thus measured, define reading. The more we break down the processes, the better we can understand and measure reading. The behavioral empiricists support the same view-

point, arguing and applying B.F. Skinner's theory to reading.[2] They say that reading is a habit system. It is based on mechanistic positive reinforcement of good habits and discouragement of bad habits. Lesson steps must be discrete. The approaches thus recommended are basal readers, workbooks in phonics, word-attack skills, vocabulary, and identification of main ideas and detail. Of course, the above follows the standardized test design. Thus we get the idea that reading is what tests measure. In our school most traditional formal teachers believe in the empiricist approach to reading.

What is needed is a definition of reading based on what children do when they read. Marks on a page are transformed into meanings in a holistic process. What students bring over a period of time to reading broadens and deepens the reading experience. Children need to be able to respect the content of reading. Children need to develop language experiences that enrich their reading. Alone and in small groups they need to know how to apply their intelligence to comprehending the meaning of experiences. The process in reading is similar to learning how to talk. All children have learned to talk. They have used trial and error and succeeded. We did not classify or categorize children before they came to school. We let them talk. We did not understand how children learned and applied their language. Yet children did learn to talk, just as they can learn to read.

What is needed is a definition of reading based on what children do when they read.

Test publishers have failed to consider problems of testing minority children. Tests are used for sorting out and determining the economic and social future of school-aged children. Problems of testing cannot be solved by recreating standardized tests for minorities based on antiquated concepts of intelligence and achievement. De Avila and Havassy[3] conclude that we need a new approach. They recommend the Program Assessment Pupil Instruction (PAPI) system. This generates two basic types of information by means of computer data-processing. The first type is statistical in nature and is meant for funding The second is for teachers: it suggests classroom activities. The test battery consists of four tests, individually administered. The first three tests are paper and pencil group tests; the fourth is individually administered. Achievement and developmental levels are considered for each child and his/her reference groups. If a child does not understand a concept, sets of classroom activities are recommended. Chronological age is used for the child's peer or reference group. However, the PAPI system can be designed so that the child's reference group can be described on the basis of grade, ethnic group, or sex, etc. I would certainly like to explore this type of testing at our school.

Our school has responded to the individual needs of students. We have broadened their school experiences. We have focused on the thinking process as well as the development of values and personalities. We have developed and participated in network activities emphasizing a humanistic mainstream climate for all of the arts, special education, bilingual, open, and traditional education. Yet we are still stuck with standardized tests for assessment of our educational quality. Fortunately, we have been able to maintain the leading position vis-a-vis other schools in our district. But the depth and quality of our innovative bundles of learning activities have created a garrison outpost

mentality. We always have to be careful to watch what we are doing. This creates a great strain on our administrative and teaching staffs.

In all our programs we attempt to develop cooperation among students in various projects. We believe students must learn to work together in real life. Standardized testing isolates the student. It destroys the efforts necessary to learn the difficult process of problem solution through cooperative efforts.

Frequent testing leads to student discouragement. Success leads to skill development through repeated performance. We must encourage and make classrooms a joyful experience for *all* children, not just the 50 percent above the standardized test level norm.

Notes

1. Robert Stake, "Position Paper: First California Conference on Educational Evaluation and Public Policy" (North Dakota Group on Evaluation Monograph, 1976), p. 52.
2. B.F. Skinner, *Beyond Freedom and Dignity* (New York: Alfred A. Knopf, 1971).
3. Edward A. De Avila and Barbara Havassy, *Some Critical Notes on Using I Q Tests For Minority Children and A Piagetean-Based Computerized Information System As An Alternative* (Stockton, California: Stockton Unified School District, 1973).

7. *Coordinating Testing, Evaluation, and Decisionmaking at the Local Level*

Blas M. Garza, Jr.
Principal, Franklin Elementary School, Santa Barbara, California

It would be safe to say that there is a very high degree of agreement in the country that the most important function of the public schools is to teach literacy. Beyond that, there are many other functions of schooling that are also considered important, but there are differing opinions as to how important. One needs only to gage the momentum of the back-to-basics movement and competency testing developments to be impressed with the relative importance that people place on literacy.

In local school districts everywhere, a great deal of time, effort, and money are spent to make the practice of teaching children how to read more effective. Learning to read and performing the function well is so important that monumental efforts have been launched by Federal and State Governments to attempt to provide direction and special assistance where local efforts need help.

The purpose of this chapter is to view decisionmaking at the local level in areas of reading and categorical programs and to describe the influence that tests and evaluations have in making these decisions and to describe the need for more coordination locally.

It should be understood that the decisionmaking process at a local school is filled with numerous distractions and pressures. One seldom has the luxury of a concentrated period of time which can be devoted to the thorough study of any one subject.

In attempting to focus attention on the decisionmaking process, I am reminded of the model which is common to legislative bodies, boards of education, and city councils. In this model, facts, figures, reports, and statistics are presented, and after some discussion, a decision which is presumably based on the evidence is rendered. While things may sometimes happen that way at an elementary school council or faculty meeting, much decisionmaking at this level does not actually follow that format.

I have pondered whether the model described above is really as clear-cut as it seems. I recall that there is a great deal of skepticism regarding research and evaluation; in essence, the suspicion is that one can prove anything with statistics. If there are such suspicions, what is the net effect that

---

Details about curriculum improvement efforts in the Franklin School and an illustrative appendix have been deleted from the full version of this chapter–*Ed.*

testing and evaluation might have on the decisionmaking process? There has been some research in this area. An article by David Cohen and Michael Garet, in summary, states the following:

> Characteristically, while lots of money has been spent on policy research, much less has been spent on assessing its consequences. A few descriptive accounts and some after-the-fact analyses, however, have shed some light on the issues. In general, efforts to improve decision making by producing better knowledge appear to have had disappointing results. Program evaluations are widely reported to have little effect on school decisions; there is similar evidence from other areas of social policy.[1]

There is a similar, though more hopeful, view in a speech by Fred Kerlinger, past president of the American Educational Research Association:

> How does research influence and change education and educational practice? The effects of research are indirect and deep and are felt only over appreciable periods of time. Deeper understanding of underlying phenomena is relatively slow, even reluctant, because it has to combat or displace fixed sets of beliefs.[2]

The insights from these two articles provided me with some reassurance regarding some of my own observations. Testing, evaluation, and research *do* have a role in decisionmaking, but it is usually combined with many other factors, including some political ones. For the active decisionmaker, all of them serve as a network of information which are carried around and applied when needed. This process is not necessarily the neatly packaged legislative body model, but it is realistic for the local administrator.

This chapter will focus on actual testing, evaluating, and management decisions at my elementary school.

The Situation

Franklin Elementary School is situated in a generally low to middle income section of Santa Barbara. It is the largest elementary school in the school district, having an average enrollment of approximately 720 students. The student population is diverse and includes children from families on welfare, families from upper levels of professional occupations, and a majority from skilled and semiskilled working families. The ethnic and racial groups represented include approximately 57 percent Spanish-surname, 15 percent black, 2 percent other minorities, and 25 percent white. Transiency is approximately 30 percent. Spanish is spoken in approximately one-third of the homes, and about 30 percent of the students are considered either of limited or non-English-speaking ability.

Because Franklin is considered a target school, there is a primary concern with reading improvement. This concern stems from findings of the district testing program. In 1970, the Santa Barbara School District began using the Metropolitan Achievement Test in reading and math, with the intention of continuing its use over a number of years. Since the State Department of Education had not been consistent in using the same test in the State testing program, it was felt that the MAT would provide local consistency for evaluation and comparative purposes. The results of these tests have been presented to the board of education every year and comparisons made on the relative performance of each school. Teachers have used the item analysis data in their planning and principals used the same information to assess program strengths and weaknesses. Public disclosure of the test scores also helped to apply pressure for improvement on the part of teachers and principals.

For a few years during the early 1970's, MAT test results were grouped according to the major racial and ethnic concentrations in the city—white, black, Spanish-surname, Asian, and other minorities. This type of analysis served to demonstrate that the composite school or district scores were not reflected evenly among the various groups. When separated, white and Asian groups reflected a much higher score than the school or district average, while black and Spanish-surnamed children scored far below. "Other minority" was a category with very few children.

On the whole, Franklin School consistently ranked among the lowest three schools in both reading and math. Grade-by-grade analysis of the test data showed that the school's first grade classes usually scored above the 50th percentile based on the publisher's norms, but that a gradual regression occurred with each succeeding grade, with the sixth grade scoring at the 26th percentile.

Beginning with the 1974-75 school year, the State of California began using the California Assessment Program, which is a matrix sampling test for first, second, third, and sixth grades. The California Assessment Program showed a pattern for the school similar to that revealed by the MAT.

> *... it is extremely difficult to explain low test scores to parents with the explanation that "We're performing as well as is expected for schools such as ours."*

Perhaps there should have been some consolation in knowing that in comparison with other schools similar to ours, our students were scoring within the expectancy band on most skill areas most of the time. That is, our students were performing as well as should be expected. However, it is extremely difficult to explain low test scores to parents with the explanation that "We're performing as well as is expected for schools such as ours."

A District Decision To Move Into Biligual Education

Because of its high concentration of Spanish-speaking students, Franklin School is deeply involved with bilingual education. This development has been slow in coming and even now needs further strengthening. However, the school district is not now employing any new teachers because of declining enrollment, and there are no more qualified personnel within.

A decision for the school district to move into bilingual education programs was made at the time the board of education was receiving Metropolitan Achievement Test scores showing low achievement of black and Spanish-surname children. As director of Intergroup Education, I attempted to impress upon members of the board of education that the reason the Spanish-surnamed children were scoring so low in reading was that many were non-English-speaking and a great many more were of limited English-speaking ability; consequently they needed help with oral language development. Since most black children were from low socioeconomic backgrounds, they too needed special oral language development.

The need for bilingual education was very controversial in 1971. However, the board of education was responsive to arguments in favor of the unique needs of the Spanish-speaking. It appropriated $10,000 to the Office of Intergroup Education for the purposes of exploring the need further and developing a plan on how a program might be implemented.

The following year, the school district was awarded a State grant for a bilingual education program. Franklin School was one of the first schools to be part of the program. Beginning with a kindergarten class, the school was to add one additional class per grade each year.

Evidence of success was slow in coming. There was much enthusiasm on the part of the participating staff and the parents, and it was readily noted that children in the program were happier and learning more. Tangible results, however, did not begin to show until the initial kindergarten class reached the fourth grade. A decision was made at this point not only to expand the program to the fifth grade, but also to add to the program a supplemental kindergarten and first grade class.

A tandem effort to improve reading achievement was the adoption of a new reading textbook for all grades. School personnel have generally felt that the reading program was weak, since there was a wide assortment of books and materials in use but no common direction.

While the school was busily exploring new ideas to improve reading, a seemingly contradictory event occurred: a decision was made in the school district not to accept any new categorical programs. The decision was not anti-improvement as such, but it did underscore teacher resentment of regulations, controls, and duplications in categorical programs. The final decision in this matter was actually made by the board of education, but it was prompted by strong teacher opposition to the introduction of any new categorical programs.

The question in this case was whether the school should make an application to the State for early childhood moneys. The State of California had made provisions through Assembly Bill 65 for expanding its efforts in early childhood education. One of the features of the new funds was that they could be spent on all children regardless of test scores, language, or poverty factors. There would be no restriction on the type of activity for which money could be spent, as long as it was designed to meet proven needs which were approved by a school site council comprised of 50 percent school staff and 50 percent parents.

With this much "pie-in-the-sky" how could teachers be opposed to the new categorical program? It seemed that teacher discontent was focused on fear of new controls and regulations in spite of promises of local control. (Two years earlier, at the request of State monitoring officials, the school had been asked to adopt three management "continuums" in reading, language, and math. The criterion-referenced systems on these subjects required a vast amount of paperwork—pre-test, teach, and post-test on every skill. Each management system contained approximately 75 distinct skills, and all information regarding the continuums had to be properly recorded in student profile folders and in class progress charts.)

It was also widely known that the teachers in the bilingual program, although very dedicated to the program and the needs of their students, were required to do much more evaluation than the continuums. They had to test for language dominance (San Diego Test of Language Proficiency), for insights into the use of language (Bilingual Syntax Measure in English and Spanish), for determining reading comprehension (Cloze Test in English and Spanish), for determining math skill development in Spanish (California Achievement Test), and for reading progress in Spanish (Santillana Criterion Referenced Management System). They also had to keep inventories on children's self-image and social development.

It is recognized that good teaching demands careful roadmaps to learning be followed, and all the instruments mentioned above are useful. The rub, though, as far as the staff was concerned, seemed to be that many instruments were deemed to be imposed without due consideration of whether the additional requirements would detract from the existing curriculum or how they would com-

plement it. Although local instrument selection studies were made by teams of local teachers, the choice was not whether to get them, but which ones to get. This left resentment among the staff, and the resentment grew as problems of program implementation developed.

Problems With Evaluation and Coordination

It was in this setting and mood that the school was considering whether to undertake a new categorical program in early childhood education. Among the information made available on A.B. 65 to teachers and parents was a question-and-answer publication in which a State official interpreted the meaning of the A.B. 65 legislation. One of the questions was: "Will local districts do the evaluation of student achievement or will the state? If the state, what indices will be used?" The official's response was:

> A.B. 65 really provides for four levels of evaluation—school level, district, State and an independent evaluation. There are requirements for different kinds of evaluation at each of these different levels, including information on student achievement and student cognitive and affective growth as well as process kinds of information about what's really happened at the school as a result of the funding that school has received.[3]

The evaluation burdens suggested by the response were an additional concern of the teachers when considering the new categorical program. Several informational bulletins were sent to parents prior to the final vote of the Franklin School site council. One of the bulletins contained a list of what I considered to be the advantages and merits of A.B. 65. It also contained a list of disadvantages drawn up by teachers, consisting of the following:

- State moneys bring State control.
- Teachers fear additional continuums and management systems.
- Teachers fear too many people bring fragmentation of classroom programs.
- The school site council involves too many meetings and time spent not in the classroom.
- More money may not provide better education.
- Planned provisions for uniformity can decrease teacher flexibility.

When the school site council met to consider whether it wished the school to be included in the appropriate section of A.B. 65, it was clear that teachers harbored a deep mistrust. The vote was eight to four in favor of inclusion, but the four negative votes were cast by the teachers!

As principal, I saw this situation as unworkable. Although the vote favored inclusion and the council could make that recommendation to the board of education, any attempt of parents and teachers working together to plan school improvements would be jeopardized by teacher resentment of A.B. 65 being forced on them—in this case by parents, principal, and other staff members.

In presenting the positive recommendation of the school site council to the board of education, the school and district administrators nonetheless recommended that the Franklin School be excluded from consideration. The board of education agreed, indicating that it wanted the council to reconsider the matter within a year.

It is appropriate to look closer at some of the conflicting factors that brought about this decision. This can bring more into focus some of the coordination problems at a local school.

Franklin School receives categorical assistance from five sources. Two of them, Federal Title I and State Educationally Disadvantaged Youth, are combined as one program with common application procedures through the State, and they also have a common evaluation procedure. Title VII,

Bilingual Education, is a Federal project with separate guidelines and evaluation procedures; it is funded directly from Washington D.C. The Emergency School Aid Act project (ESAA) is still another Federal project with its own set of guidelines and evaluation procedures and it, too, is funded directly from Washington D.C. Migrant Education is another project with separate administration, funding, and evaluation.

It frequently happens that many of our same children, by virtue of qualifying under different program guidelines, are participants in all programs and that separate evaluators analyze the same set of test data for the different programs.

This question about identical children qualifying for help from different programs raises another problem. If the same child is receiving services from four different programs, how can four different evaluators apportion for their respective programs the gains made by the student they have in common? Or, being separate and; independent programs, do they all claim full credit for progress made? Who takes the blame for failure?

It is permissible under California consolidated application guidelines to apply Title I and Educationally Disadvantaged Youth funds to unmet needs of Spanish-speaking children even if they are in the Federal Title VII Bilingual project. Help is therefore provided in the way of money for educational materials and additional instructional aides. Some of these same children also qualify for the Migrant Education program, in which case they receive special medical assistance as well as additional instructional aide help. Special assistance is provided from migrant funds for family liaison and social problems. If the children are in the Bilingual Title VII program, they also have the services of a home liaison worker.

Categorical projects, because they come from distinct State and Federal grants, have a tendency to create independent departments at the local level.

Because of the number of children needing bilingual education, two additional classes (not Title VII) were created. These classes have been funded by Title I and State EDY moneys, although they get supervision, guidance, and direction from the Title VII program. Finally, almost all of these projects have components for intergroup or human relations, but they all depend on the ESAA-sponsored Multi-Cultural Teacher Resource Center to provide for those needs.

That is perhaps sufficient to show that there is a splintering and duplication of effort locally. The categorical projects are all dedicated to the solution of problems, but together they amount to a confusing picture—a patchwork of effort. Because all projects have their own district level coordinators or directors, decisionmaking in the school reflects the results of combined decisions made at many levels. Often this means relegating the principal's role to simply accepting and coordinating directives.

Categorical projects, because they come from distinct State and Federal grants, have a tendency to create independent departments at the local level. This further serves to splinter efforts for a school principal who must deal with various directors and coordinators. Even if projects are consolidated under common administration at the school district level, local officials still have to contend with separate and distinct regulations issued from separate State and Federal agencies.

It is apparent now that State and Federal involvement in local education is here to stay and that categorical programs need to be planned more as integral parts of the total local educational effort rather than as extra appendages which are here today and gone tomorrow.

Some Concluding Observations

This chapter has presented problems and dilemmas in efforts to improve literacy, expecially for those with language problems, at one elementary school. Some observations are now made to suggest how research and development may help such a school.

1. Pressures for improvement have been great at Franklin School and it has been agreed that improvement was needed. However, the reporting practices of the tests used there and of most other normed tests are detrimental to the efforts of schools such as Franklin. At best, normed results only serve to create much anxiety among teachers, administrators, and the general public. Far better would be a system which begins with a careful assessment of school entry skills and weaknesses and relates all reporting procedures to the measurement of growth only.

   A professor once explained this problem with great clarity: "The problem with testing is that 100 percent of the parents expect 90 percent of their children to be in the top 10 percent of the class." The reasoning can be extended to school communities. All school communities want their schools to be at least above the 50th percentile. This is a vicious cycle! Will anyone above the 50th percentile ever stand still so others can catch up? A new testing frame of reference and reporting practice needs to be developed.

2. The case of overlapping services to identical children has been presented and questions have been raised regarding evaluation accountability for each project that delivers such services. Because the projects have reporting obligations to different State or Federal sources, it is unlikely that evaluation analysts at one level will know details about evaluations of the children's progress being reported elsewhere.

   Every program is attempting to remedy a specific piece of the problem, but when there is more than one grant, the pieces that each one is attempting to remedy often amount to more help than one child can use. Separately, grants can amount to more specific help than is needed in any one area; the Federal and State officials may not be able to visualize the total delivery of the services at the local school level.

   Ideally, the hand of the school principal should be strengthened to allow much more flexibility than now exists. This is not making a case against categorical funds, for it is recognized that many reforms came about because categorical moneys were created for specific purposes. It is a case, however, for examining the splintering effect in individual schools of diverse grants from State and Federal levels, and coordinating them better in order to avoid duplication of services.

3. For the sake of efficiency, all project evaluations should be concentrated in the school district's office of testing and evaluation. Financially, it is no problem to take a portion of money from each categorical project and allocate it to the district's office of testing and evaluation. The amount could be prorated depending on the size of the project. The gain is that the same people would handle all test dates and would be thoroughly familiar with them, facilitating assessments which must be made. These people would also have an overall perspective on the requirements and needs of all programs. Such a move would

save needless duplication and waste of time and energy. The evaluation consolidation procedure should be duplicated at the State and Federal levels.

4. The overall problem can be corrected by consolidating all categorical grant resources. It is a truism that grants funded through different Federal or State offices have a tendency to establish separate departments at the local level. The Federal office can see to it that all categorical grants are funded to the local school districts through the State. The State can then see to it that all Federal and State funds are delivered to the local level as a combined and integrated package. All categorical grants would therefore have a common evaluation and a common accounting system. At the local district level, all programs would then have a common administration and evaluation.

Notes

1. David K. Cohen and Michael S. Garet, "Reforming Educational Policy with Applied Social Research," *Harvard Educational Review* 45 (February 1975): 19.
2. Fred N. Kerlinger, "The Influence of Research on Education Practice," *Educational Researcher* 6 (September 1977): 8.
3. Association of California School Administrators, "Implementing A.B. 65: Questions," *Special Report* 7 (January 1978). (Responses were by Tish Bussell, Chief, Office of Governmental Affairs, State Department of Education.)

8. *Evaluation of a Planned Change Effort*

M. Claradine Johnson, Assistant Professor, Department of Personnel Services, College of Education, Wichita State University

Problems and Limitations of Evaluative Processes

During a time when financing, accountability, and credibility are critical issues in education, when their lack is literally eliminating programs and shortening school terms because of diminished public support, when charges of nonperformance are being leveled at educators because of declining test scores, when a lack of salary increases is the will of the taxpaying public, school people should be clamoring for assistance in developing evaluative plans by which they can document their professional worth. The much maligned process of evaluation is a potential lifeline for public education, if only it could be refined and accepted.

Some of those who are contributors to the field seem to agree that evaluation as a process is, indeed, "ill." In the early 1970's recognized authorities Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman, and Provus collaborated as members of a National Study Committee on Evaluation. The goals of the committee were to communicate the conceptual and methodological problems and needs of the field and to review procedures and techniques which currently could be employed.[1] The authors concluded early in their writing that evaluation was not then prepared to respond to the issues raised in the schools of the seventies.

The purpose of this paper is to acknowledge the problems of evaluation that emerged in an inner-city high school. It is hoped that this discourse will contribute to the development of some prescriptions for the "illness" that seems to plague educational evaluation not only in this local setting but also in general.

The focus is an effort at overall school renewal undertaken in a Wichita, Kansas, high school. As an administrator, I was aware that the organizational development process would need evaulation, and limited efforts were made to accomplish the task. But I did not have the sophistication, time, or school district technical and financial assistance to measure adequately the 4-year plan to revive the school. The troublesome consequences of this limited evaluation are described here.

Claradine Johnson was principal of the Wichita, Kansas, high school discussed herein when this chapter was written. She is now an assistant professor in the College of Education, Wichita State University. A review of related literature on evaluation practice has been deleted from the full version of this chapter–*Ed.*

The Setting for Planned Change

Wichita High School East is an inner-city school in that Kansas city of 260,000. In 1974 the school was typical of the city's other five high schools in its per-pupil budget allocation and staffing ratio. It was atypical in that its location is on the main drag-street adjacent to the ghetto, and its programs were housed in a 56-year-old facility. East High campus had been the locus of many racial disturbances. The press had been less than generous. When negative issues of youth were to be investigated, reporters turned to the school that was perceived to be the worst of them all for the latest story. However, there had been a time, even as late as the early sixties, when the school enjoyed tremendous prestige. The East High alumni roster boasts many local greats, and even one person who is illustrious in a wider circle, Jim Ryun.

In 1974 the student enrollment of 2,300 comprised a broad socioeconomic spread from welfare to wealth and a racial mix of 23 percent minority students, most of whom were black. Twenty-one buses arrived each morning to bring enrollees from the Assigned Attendance Area (Wichita's approach to desegregation), as well as those who "qualified" for a ride by living more than 2½ miles from school. Often there were problems either on a bus or problems later in the day that originated on the bus.

The faculty numbered approximately 100. The membership supported a pervasive internal power structure which gained its strength from the need to survive. At least 50 percent of the staff had been assigned to the school for over 5 years. There were scattered examples of quality teaching but there was little evidence of professional interest in program or staff development; nor did it seem that staff members were even interested in each other.

Problems were commonplace. Security guards and administrators patrolled the halls while faculty members carefully hid themselves in classrooms or the many nooks and crannies that had become informal lounges in the cavernous three-story building. Students were chronically absent, either by the hour or by the day. Parents had been known to pull every possible political string to have their child transferred to another high school. A computer summary of the 1974-75 school year indicated that 33 percent of the sophomore class had been withdrawn before the year was over, or had failed one or more classes.

The Change Model

It was in this setting that a planned change effort was initiated in the fall of 1975. A classic organizational development model was followed in an effort to involve the entire school community in identifying problems and proposing solutions. Essentially the plan involved the following steps:

1. Establishing awareness of problems by scrutiny of hard data—absences, failures, withdrawals.
2. Assessing needs.
3. Identifying problems to be addressed.
4. Establishing priorities.
5. Charting goals:
   a. Reviewing existing programs with an emphasis on updating curricula and incorporating learning style/teaching style awareness.
   b. Finding a process by which negative student power could be redirected to become a positive force in the management of the school.
   c. Finding a process whereby teachers would be motivated to become involved with students and assume the role of student advocates.

6. Planning programs to accomplish the goals.
7. Evaluating the programs, including the general school climate on a longitudinal basis.
8. Making plans to eliminate or expand programs based on evaluation results.

A third-party evaluation was contracted to evaluate the overall school renewal program.

Planned activities became a reality. Existing programs were reviewed, with resultant expansion in some cases, deletion in others. Those changes were undertaken departmentally. The increased awareness of the problems of the school–school-leavers, nonattenders and failing students–moved some departments to establish independent study programs to be used as alternatives to the conventional classroom. The course content was monitored and evaluated by members of the departments under the direction of the department coordinator.

The district's curriculum division became interested in the efforts being made at the school and recommended that East be the site for an Experience-Based Career Education (EBCE) program to be funded for the district by the United States Office of Education. (EBCE is an alternative school program developed with NIE funding which incorporates student experiences in the community and workplace into academic credit. EBCE was an emphasis in funding under Part C of the Vocation Educational Act at that time–*Ed.*) The program was to be a 48-month project to start in 1976. The proposal also called for third-party evaluators to assess the EBCE. The evaluators for the overall program were thus commissioned to evaluate EBCE's 14 process objectives and 12 outcome objectives by the following methods:

1. The status of process objectives was determined by interviewing program personnel and examining project records.
2. A pre-test/post-test design was used to evaluate the attainment of EBCE outcome objectives associated with academic achievement, self-esteem, career orientation, and sex bias. There were 51 students in the program and 41 controls.)
3. A self-administering checklist/open-ended response form was used to collect summary impressions of the EBCE program from students, parents, and site resource personnel.
4. Three career sites were visited by evaluators to interview resource people regarding impressions of the program.

A peer leadership program was then initiated in East after an incident of racial violence that occurred during a powerline blackout. That violence underscored the need for a program to insure a safe educational environment. The assumption behind this program was that student power could help produce a positive learning climate, and that student cooperation could alleviate many problems. It was believed that the students themselves must become involved in helping to develop and maintain a nonthreatening atmosphere in the school.

The peer leadership program, designed by a local task force, included 80 students who were identified as leaders of *all* kinds on the campus: in student government, athletics, sponsored clubs, unauthorized clubs, class-cutting, hall-walking, parking lot dis-control, pushing, and hustling. All grade levels and races and both sexes were represented.

Seven goals were defined by the peer leadership local task force and accepted by the students:

1. To develop an understanding of preventive law in society.
2. To develop communication, personal interaction, and group process skills.
3. To develop an appreciation of the values of others.

4. To provide resources and guidance to peers in helping them solve problems.
5. To develop an understanding of leadership qualities and the power of peer leadership.
6. To provide an opportunity for classroom discussion of student problems.
7. To improve students' attitudes toward school.

This program received the funding from the Lilly Endowment. It, too, employed the use of the third-party evaluators. The evaluation targets were as follows:

1. Change in social indicators (i.e., increase or decrease in juvenile delinquency referrals, absence, withdrawals, physical attacks, vandalism costs, student involvement in school activities, discipline referrals).
2. Student performance indicators.
3. Judicial review (an "adversarial" legal process of evaluation in a courtroom setting in which the program was put "on trial").

A teacher/student advisory program was piloted during the spring semester of 1976. "Home-built" questionnaires evaluated attitudes of students, parents, and teachers. The program was expanded so that in 2 additional years all students had a teacher-advisor, and 90 percent of the teaching staff was involved.

In 1975 when the overall change model was initiated, the Staff Development Division of the District provided financial support for third-party assistance in the formulation of the plan and supported the building administrator in a climate survey in the fall of 1975, spring of 1976, winter of 1977, and spring of 1978.

A district data base also documented the reductions in failures, numbers of school-leavers, and vandalism costs, and the improved attendance. Internally, increased participation in school activities and greater attendance at school events were recorded. No record was made of the decline in numbers of parental complaints or requests for transfer. The local press featured various program efforts with a positive approach.

By the school year 1977-78, significant people in the school community were saying that the school had been "turned around." Teachers had become a unit that tended to work as a joint force with administrators and volunteers; they assumed curricular leadership not only in their own departments but in district endeavors. Students were registering increased pride in the leadership abilities they were able to demonstrate among other school groups in town. Parents made supportive calls and reinforced teacher-advisors when contacts were made with the homes. Central office personnel and school board members consistently referred to "what's happening at East" as a highly desirable educational condition.

Problems of the Evaluation Process

Notwithstanding successes that occurred, from the beginning there were problems that militated against the full use of the evaluation process. Those problems were inherent in the setting. The recapitulation of the evaluation that follows is not intended to be an indictment of a school system or of individuals employed therein. It is made with the awareness that the local circumstances described are not unique. The overriding constraints to full evaluation use were perceived as follows:

1. The philosophical and theoretical position of central administration. The philosophical and theoretical position of central office administrators with regard to evaluation remained

undefined, but experience indicated there was not an extensive proactive use of evaluative services. As with most school districts, the research orientation was toward summative data required to obtain or support funding from various sources. A practical framework was lacking within which a planned change operation could be usefully evaluated.

The research division was limited in scope to meeting immediate and pressing needs of report writing and negotiations. There were no personnel available for consultation to individual schools on a regular basis. Very little money was budgeted for research and development of evaluative designs. While requests from outsiders to conduct evaluations of one sort or another for their own purposes were cleared through the research division, often there was confusion that ended up with competition for data within a single building.

District financial support in terms of released time for staff, evaluation supplies, data processing services, or third-party evaluators was in short supply. Attempts made at the building level to engage computer services, even to get a set of answer sheets scored, were futile and extremely frustrating.

2. Lack of sophistication on the part of the building administrator in the field of evaluation. As building administrator, I had had a great deal of "experience" with evaluation, but I lacked the expertise to create and propose an adequate evaluation design. Too, while I supported and promoted the concept of evaluation, my administrative thrust, like most, was toward maintenance and operation of the building. Given an adequate evaluation model there still would have been a scarcity of money and people to carry out a very extensive evaluation plan. There was a local awareness not only of what positive evaluation could do *for* the school, but also of what negative evaluation could do *to* the school. During the early days of the project, paranoia was an intermittent condition.

3. Posture of the building staff. Staff attitude toward any measurement other than each person's own classroom evaluation was another force to reckon with. The people simply were not educated as to the overriding benefits of evaluation. The process was perceived by many as a threat, a necessary evil upon occasion, an intolerable intervention, and totally irrelevant in most cases. Teachers were not willing to give time and effort to measurement. The program director of the Federally funded EBCE project would rather have had no data than face hostile faculty and students in order to collect those data necessary for the financial support of the program he was directing. The third-party evaluators experienced the same hostility in the building.

4. Student attitudes. The attitude of students reflected that of the staff: evaluation was an infringement on personal rights. They did not understand the importance of measuring the school's accomplishments. Students tended to avoid testing situations by not reporting or by not making constructive effort to accomplish the task if they were coerced into a test setting. To get a valid performance from a control group was even more difficult than from the students enrolled in a program.

5. Deficiencies in the evaluation plan and controversy over instrumentation. Many aspects of the change, as it progressed, were never measured. Ethnographic or anthropological methods that perhaps would have picked up on the social changes were not employed.

The third-party evaluators were critical of the instruments selected by the funding agency to evaluate the EBCE program because they were intended for use in a traditional controlled research setting. Such a design, given the flexibility of the program, was impossible to implement.

The judicial review used to evaluate the peer leadership program was built into the plan for political reasons: the funding agency had recommended that it be used. It was moderately successful, but less than practical in the setting. Such an evaluation required a great deal of time to prepare and sophistication about legal process on the part of those involved. Because the program participants were at all levels of academic and social skills, to implement the judicial review in a meaningful way was an extensive task. Faculty members questioned its worth in terms of time spent in preparation by both students and teachers.

The school climate measure for the overall evaluation was selected by the third-party evaluators because of its reputation as the best-known theoretically based instrument. This measure was more extensive than other climate measures available at the time. It included questionnaires for students, teachers, department coordinators, administrators, and counselors, and thus it provided information from a number of reference groups within the school. The problem with the instrument in this setting was that it was too global. Results did not give specific usable information, but the grand mean still proved to be about as useful as the 17 subscales, which were not considered to be highly reliable because the items were so few. Because appropriate data were not available from the publishers, it was impossible for the third-party evaluators to validate the internal reliability and validity locally.

The results on the measure consistently registered "no significant difference" in climate, though other social indicators throughout the school contested this. Strong inferences could not be made from the data. One value of the climate instrument was that it proved to be a good way to measure differences in perception of given issues between various groups in the school.

Unanswered Questions About Evaluation

I have attempted to express the frustration as a building administrator of not having adequate evaluation measures during a 4-year organizational development effort that was nevertheless perceived as successful. As implementation progressed there were not sufficient data to adequately make empirically based program decisions. We could not document the change to or for the local district, or promote the use of the change model to others who might be interested. It would seem that in order for evaluation in such instances to become more meaningful to administrators and districts there are several general questions that should be raised and addressed:

1. What actually constitutes an enlightened evaluation of education? Kaplan advises us that measurement is not an end in itself. The scientific worth of evaluation can be appreciated only in an instrumentalist perspective—one in which we ask what ends measurement is intended to serve, what role it is called upon to play, and what function it performs in inquiry.[2]

   Stufflebeam would have us start with a definition: "Evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives."[3] The kind of evaluation needed is determined by the type of decision to be made. Stake describes what evaluation should do as follows:

   As evaluators we should make a record of all the following: What the author or teacher or school board intends to do, what is provided in the way of an environment, the transactions between teachers and learners, the student progress, the side effects, and last and most important, the merit and shortcomings seen by persons from divergent viewpoints.[4]

Greater appreciation of the potential strengths as well as limitations of appropriate evaluation in our local school renewal program would have been helpful to the decision-makers.

2. What needs to happen for the "ills" of evaluation to be addressed? There are those who feel that first all parties must admit that there are serious problems. It has been suggested that perhaps we need a way to make it more rewarding to do evaluations.[5] Maybe if funding and technical assistance were available for practitioners who wish to design and implement plans, more progress would be made.

   Consultants presently in the field need to become more aware of their own inadequacies in preparing evaluation designs. Since these consultants are mainly university professors, some pattern of national institutes or seminars in which these problems can be discussed and some training in new approaches offered seems to be in order.[6]

   Provision of incentives to participate in evaluations as well as better technical assistance and more appropriately trained personnel would have helped in this instance.

Greater appreciation of the potential strengths as well as limitations of appropriate evaluation in our local school renewal program would have been helpful to the decisionmakers.

3. What kinds of methods are being, or can be, developed which might provide a more enlightened evaluation of education in general and of innovation or change in particular? Parlett and Hamilton describe "illuminative evaluation and the social-anthropological paradigm"[7] as an electic approach. Their primary concern is with description and interpretation rather than measurement and prediction. Such an approach—which aims to study the innovation, how it operates, how it is influenced by the setting in which it applies, what those directly concerned regard as advantages and disadvantages, and how the intellectual tasks and academic experiences of the students are affected—could have been useful locally. Further investigation is directed to discovering and documenting what it is like to be participating in the scheme, either as a pupil or as a teacher, and identifying the innovation's most significant features and critical processes.

   Egon Guba has postulated that a useful theory and practice of evaluation can be generated through the use of metaphors drawn from other fields and disciplines, such as the law and journalism. The metaphors become descriptors of circumstances and are used as a compact vehicle for transferring a great deal of information that would usually require lengthy and tedious explanations. He suggests that perhaps a set of metaphors applicable to educational evaluation could be generated and fieldtested. Hopefully such a development would aid teachers, administrators, and evaluators by providing help with communication, not only about the results of an evaluation, but also about the need for conducting an evaluation in the first place.[8] More descriptive, perhaps ethnographic, evaluative information would have helped in our program.

4. How can school administrations become aware of the value of well developed evaluation plans? In 1977 the National Association of Secondary School Principals did an in-depth study of 60 high school principals who had been identified as being effective in the principalship. Extensive questioning was done in an effort to determine the nature of an "effective principal." On the subject of evaluation the following question was asked: "How do you evaluate the outcomes of programs or projects initiated by you?"[9] The answers indicated that evaluation of programs was not high on the principals' list of priorities. The authors stated that most of the principals admitted that little, if any, systematic evaluation is done, other than what is required by others. Greater sensitization to the need for and importance of evaluation by principals would be a valuable asset in dealing with recurring programmatic change efforts of the kind described here. Perhaps universities should review their administrative training programs to put special emphasis on program evaluation.

Notes

1. Daniel L. Stufflebeam, *et al., Educational Evaluation and Decision Making* (Itasca, Illinois: F.E. Peacock Publishers, 1971).
2. Abraham Kaplan, *The Conduct of Inquiry* (New York: Chandler Publishing Co., 1964): p. 171.
3. Stufflebeam, *et al., op. cit.*, p. 139.
4. Robert E. Stake, *Language, Rationality, and Assessment* (Washington: Association for Supervision of Curriculum Development, 1969), p. 15.
5. Stufflebeam, *et al., op. cit.*, p. 345.
6. Ibid., p. 345.
7. David Hamilton *et al., Beyond the Numbers Game* (Berkley, Calif.: McCutchan, 1977), p. 19.
8. Egon G. Guba, "A Proposed Introductory Section for Metaphor Adaptation Reports." Unpublished draft, June 1978. (NIE is funding the development and test of such evaluation metaphors at the Research on Evaluation Program at the Northwest Regional Educational Laboratory, Portland, Oregon–*Ed.*)
9. Richard A. Gorton and Kenneth E. McIntyre. *The Senior High School Principalship. Volume II: The Effective Principal* (Reston, Va.: National Association of Secondary School Principals, 1978), p. 23.

Research and Evaluation Staff

9. *Responding to Conflicting Evaluation Demands*

Michael H. Kean
Executive Director, Office of Research and Evaluation, The School District of Philadelphia

Rapid, and often uncontrolled growth has characterized urban school district-based research and evaluation offices since the advent of the Elementary and Secondary Education Act of 1965. The view that evaluation, as the process of providing information for decisionmakers, is vital to effective school system administration has developed simultaneously with increasing Federal involvement in education, particularly in urban areas. Although the overall recognition and expansion of the roles of educational research and evaluation should be viewed positively, such rapid development does not take place without complications.

Offices of research and evaluation should function as service agencies which aid decisionmaking and advance instructional practices in school districts. Such offices should approach their task with a single mission—to help advance the quality of education available to children in the public schools. To reach this goal, standardized test results and demographic data must often first be considered. By providing this type of current information to a variety of clients, offices of research and evaluation help them to establish educational programs aimed at strengthening students' skills and knowledge. Such offices might also be responsible for assessing the progress of children in State and federally funded projects and evaluating programs designated as school district priorities. Findings are generally disseminated to funding agencies, project staffs, and other decisionmakers.

Educational advances can best be made when vital questions are posed and new answers are found. With experienced, knowledgeable personnel and sophisticated instruments and equipment, offices of research and evaluation have the ability to conduct studies with the needed degree of professional objectivity. This objectivity is balanced when working closely with classroom teachers, principals, and other school officials who interact daily with students. Through such collaboration, research and evaluation data acquire greater meaning. By sharing findings, the professionals in contact with schoolchildren are assisted in making curricular decisions which may modify instructional techniques.

Details of clients for evaluations and a general discussion of proposed solutions to the problems identified herein have been deleted from the full version of this chapter—*Ed.*

In short, offices of research and evaluation function as service organizations. Their *raison d'etre* is to provide specified, information-related services and products to a variety of clients.

In discussing the service orientation of offices of research and evaluation, as well as their clients and the problems which they might frequently experience in mediating between conflicting evaluation demands, it must be taken into consideration that rarely do two such offices function identically. The information presented here reflects both the author's experience as the chief administrator of one of the Nation's larger and more comprehensive research and evaluation offices, and his familiarity with other similar offices, through close contact with their directors. Though the information and suggestions which follow may be generalizable in concept, modification for use in specific organizational structures could likely be necessary and desirable in order to assure maximum utility. The information contained herein does not refer to any one particular school system.

Most offices of research and evaluation function or attempt to operate as service organizations. Virtually all such offices also function with limited, and often insufficient, resources. The lack of relationship between requested services and available resources often results in the need to negotiate which services and products are and are not available to the clients of research and evaluation offices. The need to balance client service requests due to lack of resources is but one of a larger set of conflicting evaluation demands faced by such offices.

> *... offices of research and evaluation function as service organizations ... to provide specified, information-related services and products to a variety of clients.*

In the pages which follow, the problem of responding to conflicting evaluation demands will be considered in detail. An analysis of its principal components will be undertaken by examining the clients responsible for creating evaluation demands and by identifying many of the demands most commonly made. The concluding section will contain suggestions for further research on the topic and will propose expanding the role of evaluation.

Conflicting Evaluation Demands

Evaluation is the process of providing information upon which decisions may be made. Though barely a dozen years have passed since evaluation became a required activity in major federally funded educational legislation, both evaluators and their clients, those whose data needs are to be served, have become increasingly sophisticated. Evaluation methodology has made rapid technical gains and has become more systematic and more manageable. At the same time, decisionmakers have also become increasingly sensitive to the importance of evaluative information, and as they have come to understand its value, they have begun to increasingly request more of it.

The tendency of local education agencies to provide for their own research and evaluation needs is a rapidly expanding phenomenon. This is borne out by a 3-year NIE-funded project to identify and analyze educational evaluation activities at the local level, conducted by the Center for the Study of Evaluation (CSE) at UCLA.

One of the project's purposes is to provide accurate demographic data about LEA evaluation units. In early 1978, a survey was conducted to identify school districts with organizational units having responsibility for program evaluation.

Introductory letters explaining the study were sent to all school districts (750) in the United States with enrollments greater than 10,000 students; in addition, letters were sent to a 50 percent sample (573) of districts with 5,000 to 9,999 students. Superintendents were asked to return a three-question postcard, indicating whether their districts had evaluation units.

Through the survey and followup, information was obtained about all 750 districts with enrollments of 10,000 or more; the study has also received responses from 464 of the smaller districts (81 percent). Ninety percent of the metropolitan districts (45,000 or more pupils) report having an evaluation unit, and two-thirds of the large districts (25,000 to 44,999 pupils) have such an organization. One-third of the medium districts (10,000 to 24,999 pupils) report the existence of a unit responsible for evaluation, while 16 percent of the small districts (5,000 to 9,999 pupils) responding have such a unit. Overall, 336 districts with a 10,000 or more enrollment (45 percent) have a centralized unit with responsibility for program evaluation. Students in this enrollment category represent approximately half of the 44 million students enrolled in U.S. public schools.

Evaluation has now become recognized as an integral part of the educational process. It has been included in textbooks, taught in universities, and legislated into law. Decisionmakers talk about it, though most do not fully understand it. Now that in-house evaluation units have finally gained acceptance, they simply cannot fail to "deliver."

Most offices of research and evaluation profess to be service oriented, yet they have limited resources and must serve different groups of clients with often widely varying expectations. The limited capability of such offices to respond to conflicting evaluation demands is the problem upon which this chapter will focus.

There are two major components of the problem confronted by offices of research and evaluation attempting to respond to conflicting evaluation demands. The first is identifying the numerous clients to be served and hence responsible for making the demands; the second is detailing many of the variant demands most commonly made.

The Clients

The fact that offices of research and evaluation are information-oriented organizations automatically means that in order to succeed in their mission, they must be client-centered. For the purposes of identification, the range of clients has been divided into two categories—direct and indirect.

Direct clients are those individuals or groups which are served as the result of organizational intent, hierarchical relationships, and/or fiscal support. Generally speaking, they have a direct right to receive services.

Indirect clients include individuals or groups which, though they be every bit as important, or have even greater need for service than certain of the direct clients, are in a position only to request information rather than to demand it. (This does not necessarily stop them from making demands, however.) The requests of indirect clients may also frequently be forwarded through other agencies or individuals which are in themselves direct clients.

Direct Clients–The direct clients commonly served by offices of research and evaluation include the following:

- The superintendent of schools.
- Deputy superintendents, associate superintendents, and assistant superintendents.
- District superintendents.
- Principals.
- Teachers.
- Instructional supervisors.
- Project managers.
- The office of funded programs. (This office is usually responsible for securing and administering Federal funds–and sometimes special State funds–for the school system. It often provides support for virtually all of the categorical evaluation activities conducted by offices of research and evaluation.)
- Office of State Subsidies (Reimbursable programs–This office often tends to make extensive use of demographic data. Without these data, the variety of State-mandated reporting forms could not be completed, nor could special information necessary for the State Department of Education and the State Legislature be developed.)
- Program directors. (The heads of subject matter areas [e.g., reading and mathematics] should rely heavily upon test score data, special program data, and instrument development services.)

Indirect Clients–They include:

- The board of education.
- Unions.
- Parent groups.
- Individual citizens.
- Community groups.
- City government.
- Governmental agencies.
- Colleges and universities.
- Outside research groups.
- Individual researchers.
- The media.
- Students.

Clients' Conflicting Demands

Given the broad array of clients listed in the previous section, the potential for a large number and variety of conflicting demands and needs should be readily apparent. The term "conflicting" indicates being at odds or in disagreement with another point of view. It should be pointed out, therefore, that in certain instances the needs and demands may seem to represent but a single focus. The area of conflict in such instances is that the need or demand is contrary to existing research and evaluation practice. In other cases, conflicting demands between two clients or client groups may be evident. In either situation, however, offices of research and evaluation often must mediate the conflicting needs and demands in such a way as to provide the requested services, given available resources and appropriate technical considerations.

Among the more common conflicting needs and demands voiced by evaluation clients are that:

1. Evaluation results having political repercussions or negative implications should be ignored. Conversely, in those instances where positive political payoff might result, pressure is often brought to bear for overemphasizing evaluation results.
2. Only positive evaluation results should be made public if other information has the potential to affect patronage positions negatively. For example, the negative evaluation of a reading project in which heavy reliance is placed upon the use of aides might result in that project being discontinued. If the aides are community people who receive their appointments through the patronage route, such an evaluation might not be greeted enthusiastically.
3. Evaluation results should always be negative; if negative findings do not result, information need not be provided. This is an example of the "we do not need evaluators to tell us about the good things that we are doing" philosophy. This philosophy views the purpose of evaluation only to identify problems. The same concerns with objectivity as in #2 (above) are relevant here.
4. Federally funded programs should always yield positive evaluative findings, lest further funding for such programs be denied. Some interesting notions emerge here. For example, on one hand there is the "it is OK for local funds to be misspent, but not Federal dollars" approach, and on the other the assumption that "Federal money is by nature experimental," so it does not matter how it is spent.
5. Evaluation should be used as a tool for improving management and instructional practice as opposed to the proclivity to simply disregard (without malice) or ignore it.
6. Evaluation results should not be utilized in program decisions (e.g., for the modification, creations, or discontinuation of programs), but should simply be included because the State or Federal funding source requires it. There is a pervasive fear among some decision-makers that the inclusion of evaluation results may lead to projects or programs being discontinued, and that such discontinuance means that the project administrator has failed.
7. The evaluation process should be structurally linked to the instructional organization of the school district. Unfortunately, in many instances, instructional personnel are reluctant to accept such a relationship.
8. Line personnel such as principals and instructional supervisors should make use of evaluative data. Though this expectation may be logical, it is often not supported by the line personnel's superiors, who simply fail to develop the vehicle to provide evaluative information to their stalls.
9. Evaluative data should always be accessible to the public, regardless of the form they take or their stage of development. This assumption creates serious incongruencies with the Federal legislation. The common conflicting demand is that all data be held confidential. This is not acceptable either, particularly in light of the Buckley Amendment.
10. The mass media require either sensational or highly significant results. In actuality, much of the evaluative data produced usually do not fall into either category, but rather describe program implementation; or, when they report on outcomes, they reflect small incremental changes.
11. Evaluation provides instant accountability, and evaluation results should be usable in singling out unsatisfactory personnel. If this were actually an appropriate evaluative role, it is highly unlikely that many teachers would permit an evaluator to assess a program in their classroom.

Recommendations for Research and Evaluation

The purpose of this chapter has been to examine conflicting demands made upon school system-based offices of research and evaluation by various clients.

It would be appropriate to conclude by recommending areas of further research and development which might bear upon the problem. In addition, several new evaluation roles seemed to emerge during the development of this chapter. These roles and their relationship to the problem discussed will be briefly discussed.

A Research and Development Agenda

The areas recommended for further research and development have been categorized according to focus. Though many of the suggested studies relate directly to the structure and function of offices of research and evaluation, others may be linked to the topic in a more tangential fashion.

Studies of the research office and research office functions—Tasks include:

- A comprehensive survey of research and evaluation reports over a 5-year period to find commonalities of research findings.
- A study of the internal and external communications patterns of research and evaluation offices to identify and to assess efficiency of paths of information flow.
- A study of the impact on decisionmaking of selected research projects.
- Development of a comprehensive system for dissemination of research and evaluation findings into efficient use in schools
- The development of system-wide models designed to evaluate compensatory education programs.
- A study of the use of program evaluation in the decisionmaking process in urban schools.
- Development of various methods and procedures for interpreting and using test data and a comparison of the effectiveness of those methods for pupils, for parents, for teachers, and for principals.
- A comprehensive model for empirically defining and specifically cataloging appropriate comparison populations for various evaluation needs in school districts.
- Development of ways by which to individualize services to meet specific evaluation needs.
- A study of successful practices utilized by offices of research and evaluation in balancing political pressures with the provision of objective data.
- A comparative study of the organizational structures of offices of research and evaluation, including reporting relationships within the office and the school system.
- A study of the formal mission(s) of offices of research and evaluation as they relate to the actual role(s) played by such offices.
- Development of new programs for training personnel to work in school system-based offices of research and evaluation.

Studies of fiscal support for research and evaluation—Tasks include:

- A study to determine the minimal and optimal dollar amounts and percentages of budgets for research and evaluation activities.
- A study of methods of allocating research and evaluation funds across school system priorities.
- Development of new ways of securing funds for local research and evaluation activities.

- A study of the role of local, State, and Federal Government agencies in supporting school system-based research and evaluation.

Studies of administration–Tasks include:

- A study of teacher and administrative characteristics related to the effective functioning of Federal programs.
- A study of the interrelationships of local Federal program administrative personnel and school administrators.
- Measurement competencies of principals in terms of instructional management applications.
- Studies of administrative decisionmaking at school and central office levels.
- Development of effectiveness measures of principals.
- Studies of the use made of research and evaluation data by boards of education and superintendents.
- A study of the differential use made of research and evaluation data produced for management decisions as opposed to instructional decisionmaking.
- A study of the structure and processes utilized by institutional research review committees.

Studies of attitudes–Tasks include:

- Effect on attitudes toward research and evaluation of personnel in schools who have and have not been part of the experimental group in a research project.
- The development and validation of instruments which measure:
  - o attitudes towards self and others
  - o attitudes towards teachers
  - o attitudes towards schools
  - o attitudes towards learning
  - o attitudes of administrators towards measurement
  - o attitudes of teachers towards measurement.
- Development of graphic differential scales appropriate for kindergarten and beginning year one pupils.
- A study of the relationship between beginning of the year attitudes of year one pupils and end of year performance.

Studies of instrumentation–Tasks include:

- Establish regression relationships between nationally normed standardized tests.
- Effect of systematically introduced response randomness on reliability and discrimination ability of standardized test items.
- Development of content standard scores for citywide test instruments from classroom tests.
- The development and validation of an instrument designed to assess career awareness.
- Scaling studies involving various reading tests used in citywide and individual school programs.
- Retrospective longitudinal study of pupils' scores in citywide testing programs to identify gain characteristics at various performance levels and to develop equations for score predictions.

Studies of programs–Tasks include:

- Multi-regression study of effects of various programs on academic performance.
- A longitudinal study of the relationship between participation in vocational training programs and on-the-job success.

- Cost effectiveness of selected compensatory education projects.
- A comprehensive evaluation of the services provided by institutions for neglected children funded under ESEA Title 1.
- The effect of different career development models upon disadvantaged youth in terms of occupational awareness and occupational decisionmaking.
- The impact of programs funded by the Vocational Education Act of 1968 upon the total vocational education program.
- Longitudinal study of students exposed to various compensatory education programs.

Studies of students–Tasks include:

- Profiles of the low SES and middle SES droupout.
- Profiles of "good" high school attendees and "poor" high school attendees.
- Trend analyses of achievement data.
- Effects of visual and auditory variability among students on achievement levels.
- A study of Piaget's and Kohlberg's theories of moral development in relation to academic performance and pupil attitudes.
- Academic and social characteristics and performance of students involved in alternative education programs. Also, a similar study of students who applied but were not accepted.

Studies of teachers and the teaching situation–Tasks include:

- A study of the difference between effective teachers in the various Follow Through models.
- Development of effective measures of actual and perceived environment in schools.
- Studies dealing with the relationships that may exist between teacher behaviors and pupil achievement in compensatory education projects.
- The development and validation of teacher assessment devices.
- Measurement competencies of teachers in terms of classroom applications.
- Development of effective measures for teachers.

New Evaluation Roles

It should now be obvious from the foregoing evaluation research and development agenda that school-based research and evaluation offices are called upon to serve a variety of masters by providing a broad array of products and services. The primary mission of such offices should still be the provision of information upon which decisions may be made. It has become increasingly apparent, however, that offices of research and evaluation need to begin to deal with other areas, including those related to policy planning and development. Decisionmakers often require assistance in making use of data provided to them so that the best and most logical decisions will result.

> *Information for crisis management should also be considered a necessary product of a research and evaluation office.*

Specifically, two types of information should be included in the mission of larger school system research and evaluation offices. First, information for school system development is critical if school district decisionmakers are to be able to promote positive change. Knowing what we do about the change process, however, we are aware of the role crises play in it. Hence, information for crisis management should also be considered a necessary product of a research and evaluation office

Though crises cannot always be anticipated, the extent to which a system can respond to a crisis will largely be related to the availability of objective data that the crisis manager may use in resolving the situation.

Included in the provision of information for system development and crisis management are a variety of other vital research and evaluation roles. A listing of these roles might include, but not be limited to the following:

- The provision of comparative data.
- The provision of longitudinal data.
- Serving as an expert witness in court on the objectivity and/or interpretation of data.
- The settlement of questions relating to the interpretation and methodology of research studies (a form of "technical arbitration").
- The direct provision of information to clients and interest groups.
- Serving as a source for the provision and interpretation of data to the news media.
- The provision of information which can be utilized to assist decisionmakers in setting policy.

In considering concerns related to policy, a number of questions might be dealt with—some prior to the advent of a decision to fund a program, and others during the course of that program's operation and at its completion. An example of the types of questions to which research and evaluation in its new role should provide answers follows:

- *Concerns prior to funding*
  - o Are the objectives of the program important to the school district's priorities?
  - o Is the cost of the program per person (per unit) questionable/appropriate/reasonable?
  - o Are there alternative means of funding the program?
  - o Are there alternative ways of accomplishing the objectives of the program?
  - o What would the impact of maintaining the same level, expanding, or terminating the service be on the community in terms of public support of education?
- *Concerns during program operation and at the completion of the program*
  - o What is happening?
  - o Are the objectives of the program being met?
  - o What is the evidence?
  - o Does the program seem to be operating effectively?

The notion that school system-based offices of research and evaluation should now play a major role in school system development and crisis management may be more politically pragmatic than innovative. Boards of education and school administrators have increasingly come to depend upon research and evaluation data, and their level of sophistication has been demonstrated by the difficulty of the questions for which they are now demanding answers. How effectively offices of research and evaluation provide these answers is directly linked to the offices' viability.

Conflicting demands for service may be inherent to the nature of evaluation. Yet, as Webster and Stufflebeam have so correctly observed:

> The successful evaluator or evaluation department is one that manages to minimize these conflicts while producing objective, useful reports. No evaluation department . . . has prospered through either totally ignoring the wishes of their clients nor through being dominated by them.

10. *Problems of Measuring Achievement and How They Are Being Addressed in the Portland, Oregon, Public Schools*

Victor W. Doherty
Assistant Superintendent of Evaluation,
Portland public schools

Major problems of measurement identified in the Portland School District over the past 20 years include:

1. Lack of sensitivity of achievement test scores in measuring effects of experimental efforts to improve instruction and learning.
2. Inadequacies of normative measures in describing the nature and magnitude of growth in learning.
3. Absence of satisfactory language and definitions for communicating about learning outcomes both within the profession and between the profession and the public.
4. Lack of reliability of conventional wide-range standardized test scores, especially at the extremes of achievement.
5. Lack of useful measurement tools in areas of learning not profitable for commercial publishers to service.
6. Lack of useful measurement tools in subject areas where agreement on curriculum outcomes is difficult to secure within the educational community.

It is recognized that there are other problems of testing today, including political problems, but many stem from the basic deficiencies described above. Research and development in Portland over the past 9 years have addressed most of these issues.

Developments in Goal-Based Measurement in the Portland Public Schools

For more than 20 years, the Portland School District has maintained a testing program anchored in locally constructed tests and local norms using standard scores with a mean of 50 and a standard deviation of 10. Under the leadership of Dr. George Ingebo and Dr. Dean Forbes, this program has

---

Details about the Tri-County Goal Collection and Rasch scaling procedures, including appendixes, have been deleted from the full version of this chapter—*Ed.*

seemed to serve the purpose of research and evaluation in the school system better than publishers' tests with their "national" norms and nonequal interval-derived scores. It is a tribute to the leadership of measurement personnel in the district and to the superintendents under whom they served that such a program survived the pressures that constantly urged a return to politically attractive uses of standardized tests and grade-equivalent scores.

Events of the past 9 years, however, have imposed on the Portland District both the need and the opportunity to advance its program to yet another stage of development, one which we believe represents important progress in public school measurement.

Need for Better Language To Describe Learning Outcomes

Shortly after the Central Evaluation Department was created in 1970-71, it became evident that program evaluation of the type desired could not occur without well defined learning outcomes in the various courses of study. Behavioral objectives, with their extreme specificity and stated conditions of performance, did not seem to be a viable type of outcome statement for planning and evaluating instructional programs. So we set about to create a type of statement that served these purposes more effectively. The result was the "course goal," a concise, clear statement of desired learning stated at a level of generality suitable for course planning.

The Central Evaluation Department organized a three-county effort to develop this new tool for planning and evaluation. Over an 8-year period, comprehensive, carefully classified sets of "course goals" were produced in 12 fields of study.

The tricounty goal-defining effort was intended to place a resource in the hands of teachers and administrators that would permit them to select rather than create statements of desired learning. This seemed necessary since attempts of school systems throughout the country to have teachers create such statements seemed to produce results of insufficient quality for successful planning and evaluation. The 12 course-goal collections created by the tricounty cooperative effort now provide a base for planning and measurement that is comprehensive and of acceptable quality. (See "note" at the end of this chapter for an extract from the course-goal collections.)

Development of Goal-Referenced Tests

All test items developed in Portland over the past 5 years have been referenced to goals in the tricounty collections. The district now has the ability to print out item results for each goal represented in each test developed for use in the system. Basic steps followed in developing tests in Portland include: selection of goals by curriculum personnel; creating or selecting items that measure those goals; trial administration of test modules; Rasch analysis of test items; screening of items based on the Rasch analysis and on experience with the trial administration; reviewing items for sex and race bias; developing test blueprints showing item coverage of each goal to be measured; formatting and printing tests; and finally, administering the tests for their intended uses in the system. Information on the extent of goal attainment is printed for each student and for each class for teacher use.

A second development that reinforces this goal-referencing capability and opens new measurement potentials is the Rasch test scaling procedure, promoted by Dr. Benjamin Wright and others in this country. Through this procedure, an equal-interval scale is created for a given set of items based upon information about item difficulty and total test performance for the group tested. The Rasch procedure attempts to define item difficulty with the greatest precision possible on the basis of

trial item administrations. The procedure can yield information on item difficulties for any test administered to any group; it also yields an estimate of the ability of individuals and groups tested.

What advantages does this method have over conventional test norming and scaling procedures? First, it permits establishment of a scale that is independent of a norming population. Given conditions of curricular validity and good test construction, it appears that item calibrations (estimates of item difficulty) based on administering a test to 200 or more students are stable enough for practical purposes, reaching great stability at about 300 students.

A second advantage of the Rasch procedure, and one of great importance, is the ability to create item pools through the administration of a large number of different tests, linked to one another by overlapping items. By obtaining difficulty values (calibrations) of the linking or overlapping items, and then adjusting the calibrations from one test to the other through linking constants, it is possible to place all items in all tests on a difficulty continuum. The scale thus created makes it possible to secure performance estimates that can be compared for various groups attempting any items from the pool.

To understand the importance of this procedure it is necessary to return to our goal-based system of test construction. One of the persistent objections raised by teachers to measurement, and especially to use of standardized tests, is the difficulty of finding or constructing tests that correspond to the outcomes sought by particular teachers. That objection can be overcome by a system that (1) permits teachers or school systems to select the goals they wish to have measured, (2) has pools of items that are referenced to those goals and which have been previously calibrated so that when they are administered, total score estimates can be derived that are statistically comparable to those derived from any other set of items administered from the same pools.

The combined goal-referencing and Rasch scaling capabilities appear to meet teacher needs.

Having such a large pool of calibrated items not only makes it possible to secure comparable measures for different groups working on different goals, it also makes possible the administration of simple tests to less able students and more difficult tests to more able students while retaining score-comparing and score-averaging capabilities. The various test publisher efforts to produce continuous scales through statistical manipulation of normative data cannot compete in accuracy and statistical reliability with the continuous equal-interval scale of achievement produced by Rasch analysis combined with the linking methodology developed by Dr. George Ingebo of the Portland staff.

Portland's test development work of the past several years has made increasing use of the capabilities just described.

Some Obstacles to Goal-Based Planning and Measurement and Suggestions for Overcoming Them

The preceding discussion of developments in the Portland program emphasizes the importance of using well defined outcomes (goals or objectives) and appropriate processes of selecting them within a school system as a prerequisite to valid measurement. It is necessary to understand, however, that there are problems in implementing goal- or objective-based planning and evaluation in public schools which must be addressed if this otherwise rational approach is to succeed. The conditions required are by no means easily attained. Goal-based measurement itself can help to bring about

these conditions by focusing the attention of teachers and instructional support personnel on the goals measured and on the importance of teaching directly for their attainment. But more is required.

It is easy to yield to the illusion that if only we were capable of defining learning outcomes in clear, simple language, finding methods and materials to achieve those outcomes would be a logical, uncomplicated procedure. Those who have tried it, however, know this is not true for a number of reasons.

First, in most school districts there is no organized support system to help teachers implement the goals and objectives they are required to write. It is difficult to appreciate the degree to which teachers rely on instructional materials that give readymade, day-to-day support to their instructional planning and work with students. While reliance on such materials is not looked on as creative or even as necessarily good instructional practice, teachers are faced with the need to deal every day with large numbers of children and to provide them with many different kinds of instruction. Time and logistics prohibit extensive planning of every day's activities by the teacher. Without textbooks and supplementary materials and learning systems that provide routines of learning that may be followed from day to day, none but the most gifted teacher is able to create the many activities and materials required to carry out coordinated, consistent programs of instruction that cover goals and objectives in every subject. The solution to this problem is not simple. At a minimum it will entail the development of improved models and procedures for teachers that demonstrate how to move from a selected set of goals to the design of learning experiences likely to attain them, and the development of materials that will support this type of planning.

The upshot of all this is that changing to goal-based planning and evaluation must be viewed as a difficult, long-range process.

To move wholesale into a system that requires individual teacher identification of all elements of learning as well as the development of methods and materials to meet these identified goals is a quantum leap in responsibility for which teachers are totally unprepared by prior training or by existing forms of organizational support. Failure to understand this and to deal with it appropriately can lead to disastrous consequences in a school system. These consequences include: (a) the possibility of inept identification of learning outcomes that are even less satisfactory than those that might be covered in adopted materials; (b) possible failure of teachers to find organizing principles that are appropriate, relevant, and challenging to students in meeting the goals identified; (c) inability to locate appropriate materials or to select effective methodologies for reaching these goals; and finally, (d) disenchantment with the leadership of school systems that places demands on teachers without providing the support needed to meet these demands.

The upshot of all of this is that changing to goal-based planning and evaluation must be viewed as a difficult, long-range process, and that school districts must work in this direction with appropriate regard to inservice education and development of instructional support components if the system is to succeed.

I should like to speak further about two kinds of changes needed in teacher education and school district operation before goal-based planning and evaluation can be expected to succeed.

Teacher Education

The current state of teacher education is almost chaotic. The student in a typical teacher education institution today might be taking a course in science methods in which he or she is taught to use new science programs, materials, or learning systems from a variety of publishing sources. These teachers might be trained in one or more of at least half a dozen different science programs in the elementary schools and another half a dozen at the high school level, some of which stress the relationship of activities to well defined objectives and others of which stress exploration without explicit mention of either process or knowledge outcomes. A similar statement could be made about the diversity of training being received by future teachers of social studies, mathematics, and language arts. In some of these programs, goals and objectives are stressed. In others they are hardly mentioned. At the same time, teachers are being trained to use published materials in some areas of instruction and to write goals and objectives in others, with no bridging of the gap. Where specific training in goals and objectives is given, most teachers are taught how to write behavioral objectives or some variation thereof. Despite the fact that behavioral objectives are, generally, inappropriate to use in instructional planning (they are more appropriately used as performance specification statements), teachers are still taught these inappropriate uses.

The problems in this area and the slowness of teacher education to move to a more enlightened use of goals and objectives in instructional planning are due in part to the inbreeding of the education profession. For quite a number of years the works of Bloom, Krathwohl, and others were regarded as the standard resources for writing goals and objectives in education. Teachers found those taxonomies difficult to use. Their classification systems simply did not square with the realities of teaching, and the amount of insight required to see useful relationships between mental processes and informational goals was simply too great for most teachers to handle.

Bloom's work was followed by that of Popham, Mager, and others, who developed and promoted the rather stylized behavioral objective format. This probably did more to set back the art of goal-based planning and evaluation than any movement that has occurred in the past 10 years. The requirement of a performance component for every such statement confused the distinction between statements of desired learning and specifications of performance required to indicate that learning has occurred.

It is not difficult to understand how teacher education attached itself to these two movements. They were by and large the only works of significance taking place in higher education that related to goals and objectives. The work done in the past 7 years in the metropolitan area of Portland, Oregon, which produced what is now known as the "Tri-County Goal Collections" has been aimed at providing a taxonomic structure that is educationally rather than psychologically oriented and a type of goal statement that is useful in educational planning.

Many new school districts in the country have text and supplementary adoptions for every major field of instruction. It appears that a minimal requirement for getting teachers used to the idea of goal-based planning is to achieve congruence between goal identification and the processes of text and supplementary adoptions. If teachers knew that materials were adopted because they related directly to instructional goals of the system, and if they were given training in relating the materials to these goals, then they would begin to take seriously the goal statements they often so laboriously and reluctantly prepare. If identifying goals not covered in text and other materials were a teacher responsibility together with finding materials or methods needed to achieve these goals, teachers would attach even more significance to goals.

Note

Tri-County Mathematics Program Goals

1. The student is able to use the symbols, elements, operations, and structure of whole numbers, integers, rational numbers, real numbers; and, as appropriate to needs and interests, complex numbers and other systems both finite and infinite.
2. The student is able to compute with accuracy and efficiency in operation with numbers and algebraic expressions.
3. The student is able to solve open sentences (equations, inequalities).
4. The student is able to use geometric definitions, postulates, and theorems.
5. The student is able to measure things which can be described by a number that compares the thing being measured to a specific unit, and to make estimates of measurements.
6. The student is able to use mathematical functions as represented by mathematical statements, graphs, and tables for the solution and graphing of problems.
7. The student is able to use principles of logic to develop a valid conclusion deductively or inductively.
8. The student is able to use the mathematics of probability and statistics.
9. The student is able to use the language and symbolism of sets, set operations, and their properties to relate topics and branches of mathematics.
10. The student is able to translate a practical problem into a mathematical sentence or model, find a solution for the model, and interpret the mathematical solution in the context of the problem.
11. The student is able to select and use support technology such as calculators, computers, and slide rules in the solution of mathematical problems, and of problems which require mathematical solutions.
12. The student knows the historical and cultural development and functions of counting, measuring, and of mathematical symbols and systems.
13. The student is able to develop skills in problem identification, analysis, organization, evaluation, application, and generalization.
14. The student values relationships of mathematical knowledge and skills to his or her increasing effectiveness in a variety of life roles.

Tri-County Mathematics Program Processes

PLANNING

System Goal:	Students will know and be able to apply mathematics appropriate to their current and future personal, occupational or educational needs.	
Program Goal:	The Student is able to use the symbols, elements, operations, and structure of whole numbers, integers, rational numbers, real numbers; and, as appropriate to needs and interests, complex numbers and other systems both finite and infinite.	To be found in the course goal collections
Course Goal:	The student is able to rename a rational number in all of the forms: fractional, decimal or percent.	
Instructional Goal:	The student is able to rename the fractions 1/5, 2/5, 3/5, 4/5 and 5/5 as decimals and percents.	

MEASUREMENT

Behavioral Objective:	Given the fractions 1/5, 2/5, 3/5, 4/5 and 5/5, the student will express in writing their equivalent decimals and percents.	Not integral part of Tri-County Project System of Goal-Based Planning
Performance Objective:	Given the fractions 1/5, 2/5, 3/5, 4/5 and 5/5, the student will write the equivalent decimal and percents with 80% accuracy.	

Note: At the classroom teacher level, the course goals must undergo a final translation into instructional goals and learning experiences. The Goal Development Project has chosen not to extend into this level of planning detail.

11. Some Problems of Evaluation in Large School Districts

Ronald E. Banks
Director of Evaluation,
Buffalo, New York, public schools

The singular character of large city school districts concerning evaluation is a subject which has developed a considerable literature over the last decade.[1] Problems involving evaluation in urban schools have also been explored by the media in every possible manner, usually with pious expressions of distaste and horror, which almost always conclude by blaming administrators and teachers for not resolving the problems which are delineated by certain evaluation data. Although such critical attention to evaluation in urban schools is not lacking in historical precedents, as Murray Levine[2] and others have shown, the amplitude and quantity of scholarly and media attention in recent years *have* served the function of bringing to a wide public the scope and intensity of urban school evaluation turmoil.

It would be fatuous to presume, in the face of this widespread washing of dirty linen, that any function of an urban school district would escape unscathed from the problems which are described in the literature and the press. However, the one area in the operation of large city school districts which has developed almost exclusively during this same period, when the problems of urban districts have been intensifying and when the scrutiny of the problems has also grown more widespread, is the evaluation function.

As the director of a large urban district's evaluation operation during the last decade, I will delineate some of the pressing problems facing such a district and their implications for evaluation activities.

Obviously, constraints of space prohibit dealing with every problem, or even most of the problems, which large urban districts have experienced in the last decade and with the effects of such problems on evaluation operations. Perhaps such a full-scale study should be undertaken along with a historical case study of a particular urban school district. I feel that such a large-scale study would be of great value, especially if written by a practitioner from the point of view of an internal participant.

Given the space and time restraints imposed by the present circumstance, it will be possible only to select three problem areas which have had more or less easily traceable effects on evaluation

---

Various illustrative details have been deleted from the full version of this chapter–*Ed.*

operations in a large school district. I will indicate what these effects were and what responses were made in the Buffalo school district; I will then offer some modest proposals for alternative strategies which could be employed in responding to these general problems.

Since it is obviously necessary to draw upon my experiences in one urban school district in this paper, it is only proper to point out that such apparently parochial examples can be generalized to other school districts, urban and otherwise, both large and small. It is my experience, based on many meetings with colleagues from a variety of school districts, that similar, if not identical, circumstances arise in all school districts and are responded to in similar ways. Although one school district's experiences are not a paradigm of every other district, every knowledgeable reader will recognize much in this account which is quite familiar if he or she is close to evaluation as it is practiced in any school district.

Evaluation and Economic Influences

Since public agencies such as public schools are supported by taxes and, in the case of schools, at least in most citizens' minds, by local real estate taxes, they perforce are under attack by those elements of the locality which are in favor of economies and lower real estate taxes.

Such attacks, aside from the general undertone of complaint about high taxes in general, are many times specifically aimed at particular aspects of the public schools' use of funds. Administration is frequently cited as a cost factor by "concerned taxpayers" groups, teachers' unions, and especially the media. (Of the media and how to deal with it much more must be and will be said later.) It is almost always stated as certain that cutting administrative costs would save school districts significant sums. Although this argument is not true, since administration is usually not defined accurately and it generally represents a very small expenditure in most school districts, it is an emotional one that unfortunately carries weight with superintendents and board members under pressure from outside agitation.

Insofar as evaluation is an operating phase of the administration of a large district, the argument concerning administration costs does prevent proper staffing of evaluation units, since the superintendent and the board are averse to the appointment of administrative personnel. Evaluation leadership must be compensated at and operate at an administrative level; therefore, approval to hire evaluation personnel is many times seriously hampered. This holds true even if the entire support for such personnel is from Federal or State sources, as is frequently the case. Because of this attitude, evaluation personnel must frequently be recruited at employment levels other than administrative ones.

This results in being unable to offer the necessary economic and status inducements to attract appropriately trained and experienced evaluation personnel. When this hesitation to add administrative personnel to the staff is combined with the problem of the scarcity of properly trained evaluation personnel, the recruitment of evaluation personnel becomes almost impossible.[3]

Aside from the evaluation staffing problem, the economic situation which exists in practically all school districts results in a very low priority being given to those evaluation efforts which are supported by local funds. In my district, there is virtually no budgetary support locally for evaluation effort or staff, and this is not an isolated or an unusual case.

One of the paramount reasons for the appalling lack of information on such vital problem areas as class size, teacher expectancy, and curriculum approaches at the district level is the failure of

school districts to financially support in-house evaluation departments. The failure to support evaluation is related in turn to the general failure to properly support urban education by local and State governments. To illustrate the extent of this lack of local support, evaluation—including the testing programs—in the Buffalo School District is locally funded at a level somewhat less than .004 percent of the total local budget. This certainly does not indicate a high priority of local level support.

Aside from the general insufficiency of local support, there do remain the Federal and State supported evaluation efforts. At a district level, these are unfortunately restricted to the evaluation of ESEA Title I programs and other such categorical programs, except for those few districts which might, from time to time, be involved in special studies carried on by Federal or State agencies.

It is interesting to note that, although the statutory requirement for evaluation of ESEA programs has been responsible since 1965 for the growth of evaluation efforts at the school district level, recent developments at the Federal and State education department level have resulted in greater funding problems for local districts in the area of evaluation.

less than .004 percent of the total local budget
... does not indicate a high priority of local level support.

The reason for this situation is to be found in two factors: first, the consistent underfunding of Title I has forced State education agencies (SEA's) which administer such programs to underfund evaluation at the local level. In New York State, the emphasis of the State education department has been to provide as much funding to the program operations and as little to the evaluation of the programs as possible, even though evaluation at the district level is a requirement for program operation. Secondly, the necessity for the collection of achievement data to meet SEA and Federal legislative demands has reduced evaluation of such programs to the most simplified forms of evaluation design.[4] The latter tendency is intimately connected with the low funding level for evaluation since such primitive evaluation procedures are also the cheapest to carry out in most cases.

Since SEA's administer all of the federally funded programs, the decisions concerning evaluation by the SEA's have a depressing effect on the funding of all evaluation efforts. If evaluation as an important element in decisionmaking is to be taken seriously, then there must be a serious effort made by the Federal government and the SEA's to promote and fund evaluation efforts at the local district level.

In order to accomplish this, the primary effort must originate with the regulations regarding evaluation of federally funded programs. The regulations could, but at present do not, stipulate that certain specific features must appear in evaluations nor, more important, do they specify that funds, aside from program funds, be made available to the local education agencies (LEA's) to carry out evaluations. Encouragement for innovative evaluation efforts should also be stipulated. The newly proposed regulations do not address this problem, either.

This regulatory impetus, if properly promoted by the Federal government, would translate at the SEA level to the encouragement of local evaluation efforts only if guidelines for the amount of

funds were also regularized. Based on a reasonable level of expenditure, Federal regulatory requirements for LEA evaluation funding should be set at 3 to 5 percent of program costs.

In order to insure participation of all school district staff in evaluation, it is essential that both USOE and NIE encourage the sharing of planning and implementation of all forms of evaluation. A proportion of the evaluation funding should be utilized to maintain ongoing involvement of administrators, teachers, parents, and students in all phases of evaluation. An integral aspect of this involvement should be the establishment and maintenance of evaluation and testing committees, inservice programs in these topics, and the development of materials and programs.

In my district the use of such participatory committees has been underway since the early days of Title I evaluation activities. Our experience is positive in every way. However, the extent of our activities has been much less than we would have liked, because of lack of funds. These sorts of activities are not presently supported by Federal regulations as interpreted and administered by SEA's.

Relationships of Evaluation to School Staff

The present relationships of teachers and the administration of large school districts to the processes of evaluation can be characterized, in general, as hostile. If it is not shared completely by all teachers, this attitude has nevertheless been made abundantly clear in the pronouncements of teacher organizations such as NEA. Negative attitudes of principals have also been expressed negatively recently in the Journal of the National Association of Elementary School Principals. It may be pointed out that this is not necessarily an indication of antipathy toward evaluation *per se*, but is rather directed toward standardized testing. However, in essence, most evaluation in school systems is based on standardized testing. Consequently, this expression of hostility toward testing may reasonably be considered to be directed against evaluation activities in general.

> *...most evaluation in school systems is based on standardized testing.*

Such groups raise several arguments against evaluation as testing. These may be subsumed into four major complaints:

1. It is culturally biased against certain kinds of children or is in other ways "not fair."
2. It is used to judge teacher performance when teachers are not responsible for the nature of the children they teach, and teachers have no control over most of the effect on pupils.
3. It interferes with time which should be devoted to instruction.
4. It does not measure "deeper" cognitive abilities and affective changes in pupils.

All these arguments must be taken seriously and strategies adopted by evaluators to overcome their force. The expression of feeling against evaluation/testing has sometimes resulted in action being taken by groups of teachers to refuse to give tests, to fail to report results, or to sabotage evaluation/testing programs.

In my district, the very strong, militant teacher union has taken little action of the hostile sort which has been reported in some districts. This may be pure luck; however, I prefer to believe that

it stems from the district administration's desire to take these complaints seriously and meet the hostility in a reasonable manner which has some positive reinforcement for individual teachers.

One strategy our district has adopted is the establishment of a committee of administrators, teachers, parents, test specialists, and, in some cases, students participating in the program which is to be evaluated. This group, which is chosen by the members' representative groups (e.g., teacher unions, parent advisory groups, etc.) reviews the available instruments, sometimes with presentations by publishers' representatives. Administration, scoring, and reporting must be taken into consideration as one of the technical responsibilities of the testing specialists on the committee. Our experience has been consistently positive with such committees, which we maintain on a permanent basis for optimum effectiveness and real sharing of responsibility.

The issue of the use of test data as a means of evaluating teacher or other staff performance must be met head on.

Although the above strategy does not end all general argument against testing and evaluation, it does serve to promote a collegiality of responsibility for choice of instruments and provide a "learning experience" for the members of the test selection committee. It also serves to some degree to assure that the issue of cultural bias can be discussed openly in direct relationship to the choice of instruments; in many cases, items or tests demonstrating such biases can be eliminated by careful scrutiny.

The issue of the use of test data as a means of evaluating teacher or other staff performance must be met head on, since there is no more baleful influence on evaluation than such an attitude among teachers and administrators.

School districts which allow such fears to gain credence are asking for a reaction which will inevitably wreck evaluation efforts and cast a terrible influence on teacher behaviors for years to come. Teachers are quite correct when they argue that such use of measurement data is abysmally unfair since neither they nor the school has control over the type of pupil they teach. Their argument that achievement tests cannot evaluate everything they are teaching is also true.

The first precaution which must be followed, in my opinion, to counter these legitimate concerns must be the administrative separation of the program evaluator function from any evaluation function concerning staff effectiveness or, in other words, teacher evaluation, as it is practiced in school districts. Although I argue elsewhere for the necessity of program evaluation to be closely involved in administrative planning, it cannot become an arm of teacher effectiveness studies or procedures. The two must remain separate and distinct branches of administration.

The increasing amount of teacher time taken up with noninstructional activities, even if that is most generously defined, must also be a concern of evaluators. Testing and other evaluation procedures *do* take time away from teachers teaching pupils. There is merit in this complaint and, again, it must be taken seriously and met on its own ground. Two procedures should therefore be followed. First, evaluation designs must be carefully devised to minimize teacher time in testing, filling out forms, and any other involvement in evaluation. Secondly, teachers must be provided as

great a payoff as possible *for* their involvement and the consequent reduction in their instructional time. Teachers should be given the results of tests as quickly as possible, along with detailed information regarding pupil standing, percentage of pupils below the level which has been determined as satisfactory, and all other information useful to the classroom specialist teacher. Item analyses are very desirable. All of this requires concern for planning and training teachers on the use of tests to promote the improvement of instruction.

In Buffalo, all these procedures have been followed, and teachers have given a great deal of cooperation to the evaluators. Of course, this demands the use of appropriate scoring equipment and data processing facilities; but such facilities are now ubiquitous and can be utilized by even the smallest school district.

The above discussion leads inevitably to the conclusion that, without a reasonable level of cooperation of teachers and administrators, the evaluation of educational programs cannot be undertaken. At best, it will be poorly undertaken with offhanded administration of instruments, indifferent completion of forms, and general apathy. At worst, outright sabotage can be anticipated.

If central administrations ever controlled large school districts, thanks to teacher contracts they no longer do; it then becomes incumbent on evaluators to elicit cooperation in every possible manner. Our experience in such a school district indicates that with much effort and tact it is possible to obtain a reasonable degree of cooperation from those persons most intimately involved in evaluation—principals and teachers in schools. However, this requires direct participation by teachers and principals and all others who play a role in evaluation in all phases of such evaluation. This is an area where the Federal agencies and specifically NIE can play a supportive role.

Evaluation and the Media

The discovery in recent years that test scores and similar evaluation data are newsworthy has led to direct confrontation of the urban school evaluation departments with the demands of the various news media. This problem has elicited a number of responses from evaluation specialists in large school districts and a number of recent demands from the media.

> *In relations with the media, an evaluation director must walk a very narrow line.*

Evaluation specialists are frequently called upon by their superintendents to "explain" test results and other evaluation findings for media representatives. Frequently, the expectation implied by the chief school officer is that the explanation given will go considerably beyond a mere recitation of facts, and will also serve to deflect any criticism from the school district if the data being "explained" are not particularly glowing. This sort of expectation has caused even the most detached evaluation expert a great deal of anxiety, Additional anxiety is generated by media expectations that explanations will be free of any fence-straddling or other such shilly-shallying concerning interpretation of the evaluation results.

Boiled down to instances, the superintendent would like the headline to read, "Scores Skyrocket 3 Points" while the media would like to state, "Scores Plummet 3 Points." The evaluation director cannot favor either headline and is then caught in a serious dilemma.

No district evaluation director can avoid this dilemma because he or she is part of the administration of the school district. In relations with the media, an evaluation director must walk a very narrow line, balancing between the necessity to interpret evaluation findings or test data accurately and the pitfalls of undue alteration to some recurring methodological concerns, which to the media appear like waffling.

It would simplify the matter if both the media and the educational establishment had more precise knowledge about evaluation methodology and what its findings can mean than they do at present. It would also be of great value if some of the claims concerning evaluation and its usefulness, especially the usefulness of test data, had not been so exaggerated.

Be that as it may, the interpretation of test results or evaluation findings is peculiarly the responsibility of the director of evaluation or the equivalent in a school district. This responsibility is additionally complicated by the natural desires of some factions in the community to denigrate a particular program or the school district and of other factions to extoll it. These desires are also found, on both sides, within the school district's own personnel.

In our experience, time spent acquainting the media with the background of the situation at hand is always helpful. In the case of the print media, chiefly newspapers, the reporters have limited time to file their stories—unless they are working on background or feature articles, in which case more time for briefing can be made available. In the case of radio or television, the short period given to any story on the air makes succinctness of primary value. However, even in the latter case, time spent on background can frequently be utilized to reduce the overly dramatic news spot to reasonable proportions.

I have always attempted to provide reporters with opportunities to write feature articles on various aspects of evaluation and testing. In learning about the complexities of the subject they not only teach their readers through the articles but learn themselves how to interpret data with care. Unfortunately, because of the rapid turnover of working reporters covering education, the "seminars" with reporters are a never-ending matter and must be continued with each new face. Hopefully, someday education will be regarded as a subject of as much complexity and seriousness as politics, with the consequence that specialists will be hired on all newspapers who will be somewhat more permanent and more learned in the complexities of evaluation. Until then, the continual "briefing" of reporters remains an essential element for the director of evaluation.

One strategy with the media which appears to be effective concerns the careful preparation of reports to the board of education concerning evaluation and testing results. Such reports, which are always made available to the press at the same time they are reported to the board, must be carefully worded in lay language and must state any precautions in the interpretation of data which may be necessary. Both of these considerations can minimize later abuse of the data in the media.

The relationship of testing and evaluation personnel in large school districts and the media was considered by the former group to be sufficiently important to have been the subject of a conference of Large City Test Directors in Vail, Colorado, in May 1973.[5]

NIE could potentially contribute to needs in this area in at least two ways:

1. Sponsoring meetings with the media and evaluation specialists from local and Federal educational agencies to explore ways of satisfying each group's needs. Such conferences should include the working press as well as editorial personnel.
2. Developing materials such as a media relations handbook for school personnel covering areas of conflict and the needs of the media and the evaluation specialist, with similar materials to be made available to the media. The NIE, as a relatively neutral party, might be much more acceptable to the media in this area than individual education agencies. In any case, no such material exists at present for the use of either group.

Notes

1. For example, the journal *Urban Education* has printed many articles and reports. Originally published by the University of Buffalo Foundation, it is now a Sage Publication.
2. A. Levine and M. Levine, "The Gary Plan." *Evaluation Quarterly* 1, 2 (May 1977).
3. "Properly trained personnel" in the sense used here are personnel who have the essential graduate training in the areas necessary for satisfactory performance in a large city school district evaluation unit. These include statistical and measurement courses at an intensive level, familiarity with optical scanners and data processing procedures in the creation of test data, and appropriate statistical treatment and reporting, as well as exposure to administration in an urban school setting.

   Generally, the supply of such personnel is limited. The consequence of this is that inadequately trained personnel must be hired and trained in those areas in which they are deficient. This makes the problem of financial support even more acute.
4. This tendency is caused in turn by the desire of legislative bodies at all levels of government to insist on achievement data in reading and mathematics *only* as a presumed measure of program effectiveness. Occasionally, writing is added to this basic list.
5. Reprints of this conference are available from Harcourt, Brace Jovanovich, co-sponsors, with the Denver Public Schools, of this conference.

12. What It Takes To Win: Factors in the Utilization of Evaluation Findings for Educational Improvement

Freda M. Holley
Director, Research and Evaluation
Austin Independent School District

My mother was one of 13 children in a family living in a small backwater Mississippi town. She made it as far as the fifth grade and knew the hard realities of the depression years firsthand. I learned the lessons of poverty as I grew up, and I know all too well what speaking a dialect does for you in a classroom. It was the teachers and books in the public schools I attended that gave me visions of another kind of life where learning could open the world to you.

I care about education. In particular, I care about compensatory education in a very personal way. Evaluation is my way of making a contribution. Evaluation findings must be used, however, to prove to me the value of my own role in improving our school systems. If evaluation did not lead to educational improvement, I would wish to find another way to participate.

Two additional case studies on local utilization of evaluations and a general frame work for viewing factors related to evaluation utilization have been deleted from the full version of this chapter–*Ed.*

Fortunately, over the 5 years I have served as head of a research and evaluation unit in a public school system, I believe I have seen the increasing impact of research and evaluation on the improvement of practice.

The first part of this chapter presents three examples of ways our school system used—or failed to use—evaluation results, and why. The second part presents a rationale for thinking of local school district research and evaluation units as a major channel through which to foster the utilization of national research and evaluation.

Examples of Utilization

Example One: Changing Time Use

Time is a most precious commodity. Recent research has tended to emphasize its importance in education. Researchers Wiley and Harnischfeger say in their summary of the research of the literature on the relationship of the quantity of schooling to achievement that:

> In terms of typical gains in achievement over a year's period we concluded that in schools where students receive 24 percent more schooling, they will increase their average gain in reading comprehension by two-thirds and their gains in mathematics and verbal skills by more than one-third. These tremendous effects indicate that the amount of schooling a child receives is a highly relevant factor for his achievement.[1]

Current research on teaching tends to suggest, at least tentatively, that the way the teacher expends time in the classroom has a strong relationship to learning. Yet some observational research suggest that rather large amounts of classroom time are lost to instruction; Hughes concludes, for example, that teachers in elementary schools may "devote 40 percent or more of their time to management routines and maintaining order or control."[2]

In local evaluations, our results suggested that efforts at individualization, team teaching, and supervision of aides or student teachers might be resulting in an increase in the amount of management and clerical duties required of classroom teachers, with a concurrent decrease in student instructional contact. It also appeared that multiple programs or new programs, until well established, had the same result. Moreover, it looked as though a concomitant effect of these things was a drop in achievement test scores.

This was part of the background from which Austin's Office of Research and Evaluation (ORE) planned and conducted a study of time use in its major compensatory program evaluations during the 1976-77 school year. Using a detailed observation system, ORE personnel designed a study that followed a total of 227 children during their entire school day. Students designated as Title I in Title I schools were observed. The same procedure was followed for students in sixth grade schools, some identified as recipients of State compensatory funds. Although the central question we set out to answer was whether students served by special programs were receiving more instructional time from compensatory programs—or perhaps less, as some staff were complaining—the results were devastating in that they confirmed the magnitude of the instructional time problem for all students. Austin students in general were receiving only about 3 hours and 45 minutes per day of instruction. Special program students were receiving about the same amount of instructional time as all others. Other time went for such things as lunch, between-class or hall time, and classroom management activities.

The results of this study were well publicized both internally and externally. Newspaper articles appeared; television coverage was heavy. ORE planned a readable brochure that went all over the district. Graphs were used to illustrate the findings. There were some intensely negative reactions to the study: many teachers were indignant, principals questioned the methodology, and school board members simply couldn't believe the results. Other teachers and administrators, however, confirmed the results as realistic.

As the study was repeated the next year, evidence seemed to indicate that steps were being taken to increase instructional time. The Director of Elementary Education in particular gave this high priority in his supervision of the elementary principals. The study had indicated that students served by multiple programs such as Title VII Bilingual and Title I were receiving less instructional time in some areas than those in only one such program. Therefore, the department in charge of compensatory programs responded in various ways such as attempting to reduce the overlap of Federal programs for individual students through overlap data provided by ORE. A local television station even suggested that the school administration was the "grinch who stole Christmas" because of an erroneous story that schools were being required to drop all holiday activities to gain more instructional time.

Fortunately, the University of Texas Research and Development Center for Teacher Education (funded by NIE—*Ed.*) had in recent years been engaged in research that produced suggestions for teachers on reducing time in management activities. Because of its ties to ORE and the district, the Center's researchers worked cooperatively with the Departments of Elementary Education and Developmental Programs in Austin to share their findings with teachers, principals, and other staff. Coordinators in the Department of Elementary Education developed a slide-tape presentation based on R&D research and used it with all elementary school staff.

When the results came in on June 30 that year, the findings were exhilarating. Instructional time can be increased. The data from the 1977-78 compensatory education evaluation time study showed rather dramatic increases in the amount of time allocated to the academic subjects. For example, Title I students received 24 minutes more instructional time daily in the basic skills/major content areas, non-Title I students in Title I schools received 35 minutes more, and those in non-Title I schools had 23 minutes more. These findings were in general replicated at the sixth grade schools where State compensatory education funds are being used. In addition, when comparisons

were made between the other sixth grade schools and two which had voluntarily extended their school day by 30 minutes as one way of increasing instructional time, it was found that State compensatory education students in 7-hour schools received substantially more instructional time in reading/language arts than did those in 6.5-hour schools, as well as more instructional time generally. At the same time, achievement in the elementary grades increased.

Regardless of whether a positive achievement effect can be traced directly to the increased time at the elementary level, the increased time itself can be valued. Even though 23 to 35 minutes per day may not seem like much of a gain over the 175-day school year, the gain in minutes would amount to 10 to 16 additional 6.5-hour days entirely devoted to instruction in the school year. To give a further feeling for the significance of this, if current Austin Independent School District payroll costs for classroom personnel alone were used to compute a comparative cost for this gain, an equivalent number of extra school days could be estimated to cost from a low of $2,142,000 to a high of $3,265,600. Not a bad payoff.

With today's tight budgets, such a contribution cannot be considered anything but spectacular. For the Office of Research and Evaluation it becomes our best example of evaluation utilization.

Why Did You Win? Factors Favoring Utilization in Example One

Considered as evaluation utilization, the positive factors in this example stand out.

First, the initial study was based on questions arising from a national body of research findings and a base of local evaluation findings. Evaluators thus enhanced the possibility that the results would be of value.

Next, dissemination procedures were extensive. The findings were emphasized in an open school board meeting and in numerous personal presentations. The preparation of a visually attractive followup brochure that went to just about everyone in the district revived the interest of the media

a little later in the year. Clear cut and easy-to-read graphics illustrating the findings were used in all dissemination.

Very importantly, research was available to the district that suggested actions practitioners might take to make improvements in response to the evaluation.

Fourth, because of the district's strong emphasis upon accountability, the staff felt a strong motivation to act upon the findings. The district has a procedure which requires the staff to study evaluation reports and tell the superintendent and school board what actions they will take as a result of such findings.[3]

Finally, there is undoubtedly some element of luck that all these many elements came together for this particular study at this particular time.

Example Two: An Early Evaluation on Community Aides

During the school years 1973-74 and 1974-75, the Office of Research and Evaluation carried out the evaluation of an ESAA pilot project designed to determine whether community aides trained to assist in reading instruction were bringing about higher reading achievement for minority students. The project was well implemented, the school staff was enthusiastic about the aides, and the staff development designed for the program appeared to be effectively delivered. Yet the reading achievement of students not only did not increase, but seemed to suffer as a result of the aides. One possible reason was that the management time required of the teacher in supervising the aides detracted from the instructional attention given students. Also, teachers sometimes performed routine clerical tasks while aides interacted with students. Finally, the language model provided by the aides was probably inadequate. We had tested the aides and found that in many cases the reading level of aides in the project was very low. At times, the aide's reading level was even lower than that of some students in the classes.

These results were announced in 1975. The regional ESAA administrators in the Office of Education effectively cut evaluation out of the budget for the subsequent year's project, maintaining that this cut was the result of a change in national policy. One can't blame us, however, for always suspecting there was a relationship between the results and the evaluation cut. At any rate, the aides remained in the 1975-76 project. For ORE, this was one of many disillusioning events. It was the end of our second year of existence as an office, and in this and in many other cases we failed to see any immediate action taken on the evaluation reports presented.

We had learned some hard lessons by the end of those 2 years. For example, we found just how futile it was to expect anyone to read thick final report volumes. We also experienced for the first time a project director's despair at being given a failing verdict on a project into which she had put her heart, without much indication of where she could turn for improvement.

Nonetheless, we didn't give up. During the subsequent year, we began to do a number of things differently. We emphasized dissemination much more. In fact, we installed a position devoted to training and dissemination in our office structure. We repeated our important findings, including those on aides, everywhere we went that entire year. We began to be less equivocal and less inclined to repeat all our design and statistical limitations on every finding. We said simply, "Community aides don't help students learn."

Luck helped us in that third year, too. We had begun our existence as an office three levels down in the administrative hierarchy. We reported to a departmental director who reported to an assistant

superintendent who reported to the superintendent. Through a series of personnel and organizational changes, we reported directly to the superintendent in that third year. We quickly found that this change made a tremendous difference in potential impact from evaluation findings.

By the spring of 1976 several events occurred. The pilot project on aides was redesigned, not because ESAA required it, but because a new project director, formerly the staff development specialist in the project, had no desire to continue a losing program. School principals also began to ask about alternatives to aides in the Title I and in the Title VII Bilingual programs. Finally, the school board voted to discontinue funding for a large number of district-funded aides in favor of other programs. To this day, actions on the aide findings continue to surface. For example, just recently the Board cut out a $50,000 allocation for aides from a $450,000 special allocation to bilingual programing.

When Did You Win? Long-Term versus Short-Term Utilization in Example Two

This second example of the utilization of evaluation findings was not as pleasant as the first example. Yet it also offers insight into some factors affecting utilization.

For one thing, these findings are in accord with other reports that an incubation period occurs with many findings before use can result.

One element of considerable importance was the increased impact due to the change in organizational status of the Office of Research and Evaluation in Austin. I saw a definite change in the way the staff responded to ORE in both of the organizational levels. More importantly, we suddenly had immediate access to the time and the attention of the district's top administrators.

There was also a change in informal status, however. The 3d year we had credibility going for us. At the end of the 2d year, in addition to the negative report on the aide program, we had also produced a study that told the district its implementation of the Individually Guided Education program was resulting in lower student achievement. We suggested that at least one major reason for that was a real lack of implementation of the program, and that this probably resulted from the failure of the school board and administration to provide the resources the program called for to the participating schools. That we were permitted to make such statements amazed many and indicated to just about everyone including the school board itself that we had the independence to make accountability judgments. By the time this sank in, we were well into our 3d year, with everyone allocating a great deal more credit to what ORE said.

The third year meant also that we had two full years of evaluation data from which to speak to the entire district not only about aides, but about numerous other topics. Information itself has a status and power value.

Finally, ORE was much better at dissemination by the 3d year. We had discovered the great gimmicks of Chartpak and Clipart. We made more speeches and more one-page information summaries. We also learned to use the media. We wrote press releases and talked to reporters.

Example Three: Required Reports for External Agencies

The Austin evaluation unit expends considerable energy completing required reports for the U.S. Office of Education and the Texas Education Agency. These reports typically summarize achievement levels, numbers of program participants, number and cost of staff, and other such details. I have yet to see any evidence of the use of any of these reports in the district. I would feel better about our energy expenditure if I thought our reports were being used to affect practice at either the State or local level. My vision is that they fill warehouses in Austin and in Washington.

Required Losing: Factors That Prevent the Use of Required Evaluation Reports at the Local Level

Since we have found in the district that a common approach and format for all evaluation facilitates communication about results, cuts down on the time required for communication, and facilitates user response, we have developed one reporting style for all reports. They are built around "decision questions" elicited from district staff and the school board prior to or during the evaluation. The required formats are not only unlike our district reports, they are different for each Federal or State program. Thus, communicating with anyone about them requires considerable time. Since time for communication is so scarce, it is more feasible for us to redesign the information in the required reports to our district style before trying to transmit it.

Most required reports emphasize lists of numbers that are in themselves devoid of meaning. Unfortunately, most agencies seem to think that such numbers are the end of evaluation. For example, a report of the number of Title I students we serve at a given level of gain and cost is meaningful only if we compare our figures to other school districts' per-pupil cost and per-pupil gain. We are rarely given such feedback information; in fact, even when we insist upon it, we have a hard time getting information about results in other school districts.

Another problem has been the emphasis upon objectives. Objectives are a good planning device, but usually a poor evaluation device. Since programs are rarely planned with adequate baseline

data, the numbers in most program proposals are pretty wild guesses. They may also be so inadequately controlled that they are developed more for the project staff protection than as realistic goals.

If the U.S. Department of Education could provide evaluation guidelines at the national level that stressed the need for implementation and process evaluation with a recognition of the cost of these activities, we might be able to structure evaluation that was more useful locally. The development of such procedures as the current "Title I Evaluation Models"—which are helpful in some senses, I'm sure, but that don't meet my definition of evaluation models at all since they are merely statistical approaches to handling test data—do little to help us with such problems. The current models also serve to reinforce for State and Federal program officers, who usually don't have the least understanding of evaluation, the idea that test data are all that should be funded in evaluation.

I always favor the best available approach to measuring program outcomes, but I think this is rarely enough to result in educational improvement from evaluation. Since, in most cases, the staff is doing everything they know how to do, the evaluation should provide some guidance on what they can or should not do in program operation. In many cases, I believe the best thing we have done to improve programs is to provide staff people with a good literature review about available options. Such activities should fall within evaluation guidelines.

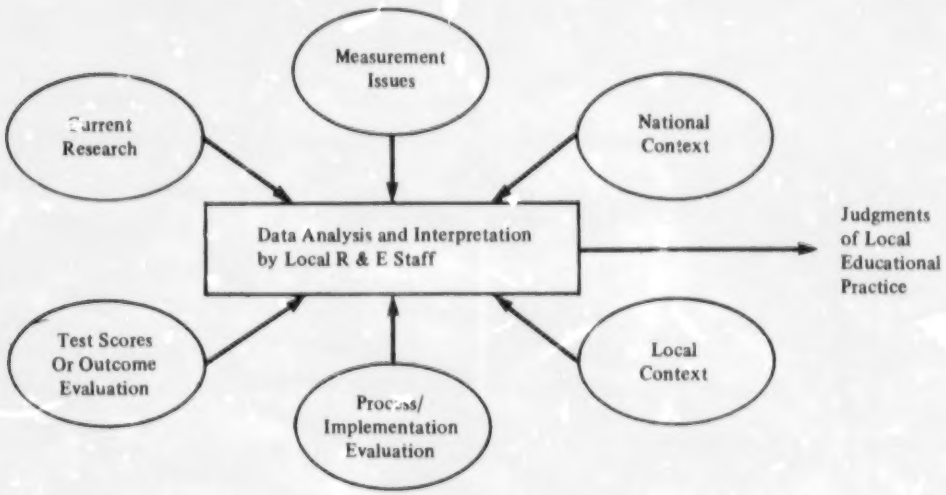
A Rationale for Fostering the Utilization of National Research and Evaluation Findings Through School System Offices of Research and Evaluation

Fifteen or 20 years ago the position of evaluator was virtually nonexistent in the public schools. In many systems there was a person responsible for "research," but the most that role generally covered was sending out a survey to other school systems in order to compare practices from one school district to another. The late sixties changed this picture dramatically. When the Elementary and Secondary Education Act was passed in 1965, with its requirements for program evaluation, the growth of the field began. One recent survey reported that in 35 large urban districts responding,

the expenditure for research, evaluation, and testing reached $33,906,888; but more importantly, those same districts spent $8,937,980,335 on general education.[4] Local/State fund expenditures went as high as $1,450,835 in one district and Federal fund expenditures as high as $10,000,000 in another for research and evaluation. Thus per-pupil expenditures for local research and evaluation ranged from $18.05 to $.80/student. Since the number of research and evaluation units is now quite large, the Center for the Study of Evaluation at UCLA (funded by NIE–*Ed.*) has conducted a study to identify just how many there are. The total number, multiplied by the expenditures we can estimate from the Webster-Stufflebeam survey, represents a considerable evaluation resource. Where such resources are properly harnessed to produce quality evaluation information tied into the national bank of evaluation information, our knowledge of educational practice increases.

I would maintain, however, that an effective local evaluation unit would be productive in a more important way. The local evaluation unit in today's school is charged with the responsibility for monitoring and determining the worth of local practices. Its ability to carry out this function depends on the extent of its ties to the national and local scene, as illustrated by the top circles in figure 1. This will affect the unit's design input from the local level and its ability to interpret it adequately and to make good judgments about practice.

Figure 1. Local Evaluation Unit Schematic



Finally, required reports usually provide little information about needed changes in activities. Most of those making evaluation allocations believe that the presence of achievement testing is the ultimate and only evaluation; anything beyond is wildly extravagant. Thus, few programs have adequate process evaluation. Reports on "nonprograms" are probably more prevalent than not. One customary staff response to such reports, therefore, is to adjust the outcome objectives down to a more "reasonable" level.

Although we have been successful locally in getting the State agency to approve more extensive evaluation in Title I, we have had a real struggle every year. It was even more intense this year than

it has ever been before, even in the light of the very real dollar payoff we could point to in the time study. Not much help comes from the national level when it promulgates "Title I models" that completely ignore the existence of process or implementation evaluation.

Where the evaluator is trained to consider and promote utilization, the eventual net effect should be improvements in local practice. Where the target of that improved practice is in the billions of dollars as indicated above, the impact should be considerable.

Of course, I am not suggesting that this is the only way to bring about educational improvement. Indeed, there are many alternative routes and no route should be considered exclusively. There is, however, a certain economy of motion to visualizing the evaluation unit in a school system as one of the essential elements for bringing about improvement. Evaluation units exist already for such purposes. Their staff members are the ones who attend such meetings as those of the American Educational Research Association, where presentations of new research and evaluation information are most likely to occur. They are the ones most likely to read research. In a communications evaluation that our office conducted in Austin, we found the most frequently read journal in the district—actually the only one read by most teachers and administrators—was the *Texas Outlook*, a publication of the Texas teacher and administrator organization. Evaluation staff, by contrast, will be far more likely to subscribe to an array of research-oriented publications. In our unit most evaluators receive the journals of AERA and APA and independent publications such as the Journal of Research and Development.

Given such pre-existing links with research and the fact that research units possess the evaluation resources and techniques in most school districts, then it makes sense to capitalize on their availability.

While other writers have effectively discussed the problems the evaluator faces in utilization, I would like to approach the problem of enhancing evaluation utilization through what I perceive as the most practical approach. That approach is to improve the capabilities of evaluators and to

provide them with additional dissemination resources. I don't believe, for example, that changing the characteristics of organizations or evaluation users is a viable approach for NIE or OE in terms of numbers and size. I do believe that enhancing the capability of evaluation units or evaluators is within reason. The evaluators could then carry the ball to other school system staff, thus extending resources.

Summary

This chapter began with some examples of evaluation use and nonuse in one public school system. It tried to analyze these case studies to see if there were identifiable factors behind utilization.

The chapter then discussed why it might be wise to foster utilization through the local research and evaluation units that already exist and are tied to the national research scene.

The picture that I hope emerges from this chapter is that local utilization is a very reasonable goal for every evaluation study conducted in the public schools and for every research and evaluation study carried out at the national level. This is a goal toward which I hope that I and my colleagues may contribute.

Notes

1. D.E. Wiley and Annegret Harnischfeger, "Explosion of a Myth: Quantity of Schooling and Exposure to Instruction, major educational vehicles" *Educational Researcher* (April 1978): 7-12.
2. M.M. Hughes, *Assessment of the Quality of Teaching in Elementary Schools* (Salt Lake City: University of Utah, 1959).
3. Freda M. Holley and Ann M. Lee, "Beyond Dissemination," paper presented at the annual meeting of the American Education Research Association in Toronto (March 1978).
4. W.J. Webster and D.L. Stufflebeam, "The State of Theory and Practice in Educational Evaluation in Large Urban School Districts," invited address presented at the annual meeting of the American Educational Research Association in Toronto (March 1978).

13. *Producing Quality Program Evaluation in Education and Using It: The Washington, D.C., Experience*

June D. Bland
Assistant For Evaluation, Division of Research and Evaluation,
The Public Schools of the District of Columbia

Use of Program Evaluation

Most often, educational program evaluation reports serve one or more of four purposes:

1. As documentation; that is, a descriptive history of the program which provides evidence that the program existed.
2. For accountability, or evidence that management objectives and responsibilities were met.
3. As evidence of the effectiveness of the program in that observed outcomes could only be attributable to the program.
4. For decisionmaking: to assist administrators in determining whether a particular program should be continued, expanded, reduced, altered, or discontinued.

Of course, while all of the preceding functions are valid components of the evaluative process, it is only when evaluation findings actually contribute to program-related decisions that the process achieves maximum utilization. There may be several reasons why program evaluations are not used as decisionmaking tools. For example:

1. There is simply a lack of understanding of what the evaluative process is, and it is seen as threatening. Unfortunately, many of the efforts which may have been intended to establish program evaluation as a legitimate endeavor have sometimes enshrouded the process with a certain mystique. The practitioner may be unclear about the research purpose and procedures involved. Thus, the combination of the mystique and fear of the unknown creates hostility. The greatest desire of the administration and staff of the program being evaluated is to "just get it over with."

Further details about the design and findings from the 1975-76 Title I evaluation discussed have been deleted from the full version of this chapter–*Ed.*

2. The evaluation process may not fit into the program planning cycle. Most programs, especially federally funded programs, plan on a 1-year cycle to correspond with funding which is renewed annually. For a program in the D.C. public school system to begin operation in the fall, purchases and logistical planning must take place in the previous spring. If the program is to be continued the following year, then the proposal for continuation must be prepared in midyear; therefore, the continuation proposal may be developed at the time in which the program is still experiencing startup difficulties. Proposals are then reviewed by several levels within the school system which have the responsibility for approval, and some proposals must go to the U.S. Office of Education or other funding agencies for further review. Even in those rare instances where evaluators have been involved since initial project implementation, there is seldom sufficient evidence or time available to process data which would provide the reviewing administrators and policymakers with information in midyear that would support a decision to continue, not to continue, or to alter the continuation of a program.
3. Perhaps there is a misunderstanding of the decisionmaking process as well, particularly by those of us who are involved with the formal evaluative process. In the formal evaluation process, researchers gather as much objective data as possible and after as thorough an analysis as time and resources permit, conclusions are drawn and recommendations are made based upon the data findings. On the other hand, administrators and policymakers are many times responding to a different set of decisionmaking information which is sometimes more personal, political, and immediate. Therefore, the evaluation may be only one of several factors which affect the final decision, and whereas the evaluation may provide the most objective information, it does not necessarily have the most influence.
4. The quality of the evaluation information provided is questionable or the evaluation information that is provided is not what is needed.

A Different Experience

In spite of the all too familiar complaint that evaluation results are not used, there are examples to the contrary. Rather than bemoaning the negative, examining and analyzing the successes may improve our capability to provide evaluation information which will be used.

At the 1978 Annual Meeting of the American Educational Research Association, a document titled *Evaluation of the ESEA Title I Program of the Public Schools of the District of Columbia, 1975-76: Final Evaluation Report*[1] received awards in two categories in the evaluation report competition sponsored by Division H: "The Best Executive Summary" and "The Most Definitive Action Taken by a Board of Education in Response to Findings of an Evaluation Report." The second award was a result of the extensive use of evaluation findings from the 1975-76 report by the program staff in developing the 1977-78 Title I program proposal which was approved by the board of education. The awards reflected both the quality of the evaluation report and the utilization of the results. D.C. Public Schools was the only local educational agency to receive two awards. In subsequent reviews by firms or agencies under contract to or associated with the U.S. Office of Education or the National Institute of Education, the quality of the report continued to generate praise.

Needless to say, those of us associated with the production of the report were very pleased: the contractors for the evaluation, the D.C. Division of Research and Evaluation, and the Title I Program office. In retrospect, those of us who were caught up in the excitement of "getting things

done" realize that there is a need to restate and review those events which contributed to the development of "a winner."

The New Evaluation Process

Under the leadership and management of the Division of Research and Evaluation, D.C. Public Schools initiated and has continued a comprehensive evaluation effort since 1966. This was contracted out to universities or private firms. Also as a part of the initial evaluation, an advisory committee, whose members were nationally recognized for their expertise in educational research and evaluation, was formed to serve as consultant to the evaluation project. Therefore, for over 10 years, external evaluations have provided an objective and in-depth analysis of program data. The data base included standardized achievement tests administered to all students in the target grades served by the program and questionnaires, scales, inventories, and other measures, both standardized and locally developed, which were administered to teachers, administrators, students, and parents. In addition, a staff consisting of three research professionals and one clerk was provided to assist the Assistant Superintendent for Research and Evaluation in managing and monitoring evaluation activities and coordinating the data collection effort associated with the external evaluation.

The results of evaluation efforts since 1966 have provided priorities for funding programs based upon performance data (there were over 50 programs) and introduced a statistical model which would provide for a continuing system for evaluating the long-range effects of individual Title I programs on a number of important aspects of pupil performance and behavior. Local norms were even established for the standardized achievement test administered for the evaluation.

Specifications for evaluations were regularly developed in consultation with program administrators. Until 1975, however, the procedure for the conduct of the evaluation was generally the traditional one in which the evaluation contractors developed the strategy based upon the program design, developed the instruments, collected and analyzed the data, and reported to the school system in formal documents and in briefing sessions to further explain and demonstrate data findings. In 1975, although the procedure for obtaining an evaluator did not change, the process for evaluating the 1974-75 Title I program did. This change occurred because the contractor selected that year utilized a procedure described as the Information Based Evaluation Model (IBE).[2] The conceptual framework of the model views supplying information to individuals in decisionmaking roles as the primary task of evaluation. Unlike objective-based evaluations,

> the model focuses on evaluation questions and the ways these questions can be answered most usefully for different audiences. Information-based evaluation recognizes that an evaluation must be dynamic if it is to be responsive. Program objectives rarely change during the project year; thus, the objectives-based evaluation is static and methodical in responding to the information requirements. Information-based evaluation accepts the fluidity of information needs and the posing of new questions throughout the program cycle. The IBE Model addresses three primary classes of information and evaluation activity: (1) product (summative) evaluation; (2) process (formative) evaluation, and (3) process/product evaluation. Product evaluation assesses program outcomes; process evaluation monitors strategies and procedures designed to change student or teacher behavior. Process/product evaluation explores the relationship between products and processes and seeks to determine which dimensions of a particular program lead to successful outcomes and how these dimensions can be replicated. Although often ignored, process/product evaluation is programmatically more important than either product or process evaluation.

> Product evaluation asks how the students or teachers are different after exposure to the new program, and process evaluation asks what strategies differentiate the Title I program from traditional approaches and whether these strategies were implemented. Process/product evaluation asks about the relationship between instructional strategies and the outcomes of the program.

Beginning in January 1975, a series of meetings designated as "evaluation design conferences" was held with the contractors and representatives from the Division of Research and Evaluation and the Title I program office. At these design conferences, a list of information needs was identified, and given priority for the 1974-75 school year:

1. Cognitive (reading and math)
2. Affective (self concept/attitude)
3. Staff development
4. Process and materials
5. Parent and community involvement
6. Management and administration
7. Supportive services
8. Communication and dissemination

In addition, the individuals who would be the most likely users of the information were also identified:

1. Title I staff
2. Superintendent/Board of Education
3. Title I schools (Parent Advisory Councils, teachers, principals)
4. Department of Federal Programs (D.C. Public Schools)
5. Catholic Office of Education (District of Columbia)

It is interesting to note that while the U.S. Office of Education required a copy of the evaluation, it was not seen as a primary user of the information. In collaboration with D.C. Title I and Division of Research and Evaluation staff, the evaluators then identified evaluation questions and data needed to respond to them. Data collection for the summative 1974-75 evaluations occurred in April 1975. (Other evaluative data had already been collected.)

During design conferences, very absorbing research questions surfaced which extended beyond the original evaluation specifications: How do Title I students differ from non-Title I students in the attainment of conservation skills (a Piagetian concept)? How does birth weight (5 pounds and below vs. larger birth weights) correlate with academic achievement? What is the relationship between teacher knowledge of the subject matter taught and student performance on achievement measures? These substudies, as they were called, were somewhat of a departure from previous evaluations. Additionally, they were not so much for the purpose of evaluating an existing program objective as they were for the purpose of exploring program-related hypotheses. Some of the substudies were reported at AERA and elsewhere.

There was a consensus among Title I program and research staff that the evaluation approach (IBE) had been very successful in 1974-75. After a review of competitive bids, the same firms were selected to conduct the evaluation of the 1975-76 Title I program. As in the previous year, a series of design conferences was held from January through March to confirm the parameters of the evaluation, but a broader representation of staff, including teachers and principals, was included. It was

anticipated that more diverse representation would result in a more responsive approach to the needs of the field staff. This would reduce apprehensions that occurred the previous year which, it was theorized, contributed to the low response rate on some instruments. Particular attention was given to the composition of the group that would provide direction in the design of teacher instrumentation. Three suggestions were made very forcefully by that group:

1. That teacher knowledge surveys not be repeated: it was felt that their use the previous year had been threatening to teachers who were not certain about how the information was to be used; also, there was not time to design and pilot another instrument that would appropriately test teachers of different grade levels.
2. That each Title I school be briefed on results from the previous year's findings by a team with representation from the evaluator, Title I program office, and the Title I research and evaluation staff.
3. That teachers not be assembled as a group for the administration of the evaluation instruments, but that the instruments be distributed to each teacher for completion at his or her convenience over a longer period of time.

As a result of these changes in data collection methodology, both the return rates and the accuracy of the data collected increased. Many teachers commented that it was the first time that evaluation findings resulting from information they had provided had been shared with them or that they had been briefed about their roles in an upcoming evaluation.

Many teachers commented that it was the first time that evaluation findings had been shared.

There was a noticeable change in the climate of the initial meetings between the contractor and the Title I program staff in the 1975-76 evaluation which was an obvious spinoff of the involvement from the previous year. This change was observed by the contractors and noted in the final evaluation report of the 1975-76 Title I program:

> An intrinsic artifact of the Information Based Evaluation (IBE) method of design is the growth in evaluation sophistication of the client. This comes about as a result of the close client-evaluator interaction, which is an integral part of the method of IBE evaluation design. As a result of the 1974-75 evaluation and the intensive effort which went into its accomplishment, the evaluators felt that the first design conference to be held under the aegis of the new contract would be excellent. We were not disappointed.[3]

Alterations and additions to the proposed evaluation strategy resulted from the design conferences. These design modifications were a response to the need for providing the optimum allocation of resources and time, and to other evaluation constraints which were identified. The contractors again observed:

> These enumerated changes, as well as others not already mentioned, result directly from an alteration in the level of expectations of staff, both Title I and the Division of Research and Evaluation. This difference in expectancy level stems partly from participation in the

design of the 1974-75 evaluations; it also derives from the realization that individual information needs can be translated into methodologics which ensure the fulfillment of those needs.[4]

In August 1976 the evaluators were invited to present their preliminary findings on the 1975-76 program at a planning conference for the development of the 1977-78 Title I program proposal. In attendance at this conference were principals, parents, teachers, Title I program staff, research and evaluation staff, board members, and selected officers from throughout the school system. The conference was conducted as a retreat over a period of several days at a facility several miles outside Washington, D.C. The presentation by the evaluator provided an opportunity for persons who were responsible for program planning to obtain evaluation information firsthand and receive clarification of any areas which could be a source of confusion. On the other hand, the evaluators had the opportunity to receive immediate feedback from an audience which represented each of the consumer groups which were to utilize the information contained in the final report. While the interaction between the evaluator and audience sometimes produced lively exchanges of viewpoints, it was a stimulating environment in which researchers and information users shared insights about the meaning of observations that each had acquired about the same conditions, but from different vantage points.

In the spring of 1977, the proposal for the 1977-78 Title I program was presented to the board of education and was approved. Almost all of the recommendations contained in the final evaluation of the 1975-76 program that related to situations over which the program had decisionmaking authority were responded to in the proposed new program. Each recommendation and resulting change was cited in the program administration's presentation to the board.

Recapitulation: How Did We Get There?

The level of enthusiasm and inquiry which characterized the 1975-76 Title I program evaluation contributed to the production of an award-winning and highly praised document. For future reference and, hopefully, replication, the essential ingredients of that experience follow.

The most notable ingredient of the 1975-76 evaluation experience was the exceptionally broad-based representation of educational professionals participating in the planning for the evaluation. With the introduction of the Informatics Based Evaluation approach, the level of interest in and the articulation about evaluation needs increased as the diversity in the membership of the planning sessions grew. Without underestimating the contribution of highly skilled technicians in the collection and manipulation of data, the contractors' introduction of a procedure which allowed for the extensive involvement of the school staff in particular was the most obvious departure from previous evaluations. As a result, many staff suggestions were made which contributed to the success of the effort. For example, all teachers who were to complete evaluation instruments were briefed on the outcomes of the previous year's evaluation and given an overview of their involvement in the upcoming evaluation. That activity alone was considered to be responsible for the significant increase in the response rate over that of the previous year. The level of response and completeness of the data enabled the evaluator to perform an in-depth analysis which had been proposed the previous year but which had had to be abandoned because of the low response on some instruments.

There were instances where the original research design was altered because of insights shared by teachers and administrators based upon their daily experience in the school environment. While there was a succession of meetings, each had a specific purpose, none was redundant, and new dimensions were thereby added to the study.

Cumulative knowledge about the evaluation process by some members of the program administration and staff was another important and necessary ingredient that contributed to the success of the 1975-76 evaluation. It can be said that from the very beginning of Title I in the District of Columbia, program personnel were introduced to sophisticated treatment of pragmatic concerns. Evaluations have been performed by trained and experienced educational researchers under contract to D.C. Public Schools. While the purpose of this chapter is not to promote contractual evaluation, in this instance the arrangement has provided a thoroughly objective assessment of a program which is often very sensitive to the politics of the school community. Technical competence in evaluation designs employed and annual briefings on findings by the contractors have contributed to the growth and development of the D.C. program staff in evaluation methodology.

The onsite participation of the 1975-76 evaluators in the planning of the Title I program for the 1977-78 year was an unanticipated bonus. The schedule of the Title I program enabled persons with responsibilities for planning and developing the program proposal to obtain feedback directly from the researchers conducting the previous year's evaluation. By the time the planning conference was held in late summer of 1976, all of the data from the preceding year had been collected, processed, and analyzed, and preliminary findings were available. This information was fed into the planning process at the very beginning of the planning cycle and additional information was provided as it became available or was needed. This timely sharing of evaluation findings contributed to the high rate of utilization of the evaluation recommendations. That these findings were also in response to many of the questions raised by the program staff was another major factor which contributed to the utilization. Had no information been shared until the completion of the final document (February 1977, in this case) the proposal would have been already completed without the benefit of that assessment.

The importance of interpersonal relations, ... should not be underestimated.

The evaluator also benefited from the sharing of findings before the completion of the final report. Information which was not clear or was ambiguous was clarified, verified, or further analyzed. In other words, it was possible to test the utility of much of the information before the final printing of the document.

A final but no less important ingredient to the success of the 1975-76 evaluation effort was the sensitivity of the evaluators. In addition to the technical expertise which they brought to the task, they were experienced in working with both educational professionals and laypersons, with and without knowledge of research procedures. There were many occasions when the evaluators demonstrated that their abiding interest was in providing meaningful information to the users of that information and using their technical expertise to ensure the correctness of the findings, not to prepare a scholarly document that could obscure the meaning of the outcomes. The importance of interpersonal relations, seldom mentioned in evaluation models or presentations of results, should not be underestimated.

Of course, as in most endeavors, there are some precautions to be observed in conducting the type of evaluation described in this chapter. The amount of involvement experienced in this par-

ticular evaluation requires time and staff commitment to coordinate, schedule, notify, explain, review, brief, and teach. Once the evaluation design was finalized (in approximately three sessions, including two that lasted a full day with staff representing the Division of Research and Evaluation and the Title I program office), several committees were formed to guide the development and/or selection of instrumentation. The number of meetings for this phase ranged from two to five. The responsibility of members of the committees was to review the instrumentation and make suggestions for changes both in content and in administration procedures. Such involvement in this process was to insure that the evaluation content was relevant and its implementation feasible. The time required for this effort must be weighed against other responsibilities program staff may have to put aside temporarily. In addition, the involvement of Division of Research and Evaluation staff in the Title I evaluation is not limited to just those individuals funded by Title I. Over the years, a tremendous number of person-hours have been expended by staff supported by non-Federal school system funds on federally required evaluations. This effort has contributed to the development of quality products.

While it is not necessary for a quality product to have an exorbitant price tag, educational administrators assume too often that evaluations can be conducted at no cost. An evaluation effort will always require an expenditure of resources, whether in the form of time to perform the tasks required or purchasing the services of professionals to perform the evaluation tasks. The larger the scope of the program and the more involved the analysis, the greater the expertise needed and the higher the cost. The quality and scope of the evaluations which preceded that of the 1975-76 program in all probability contributed to the willingness and readiness of the staff to participate in the latter.

Program planning cycles and program evaluation activity must be better synchronized. Many program planning cycles require that the evaluation information from the current year serve as input into the succeeding year. In many large systems and large programs, however, it is almost impossible to provide such quick turnaround in data unless data collection terminates in midyear. Of course, early termination may result in incomplete or misleading information. It would be more realistic to schedule the utilization of findings from "year one" of a program as the basis for planning "year three." This was how our 1975-76 evaluation was used to plan for the 1977-78 Title I program. Where possible and necessary, activities in "year two" which are inconsistent with the results from "year one" could be adjusted.

Finally, researchers, whether inside or outside of the system, must be extremely sensitive to the needs as well as the limitations of program staff relative to research theory and application. The training and experience of practitioners do not contribute to their understanding of research procedure. Therefore, researchers would be wise to provide gentle guidance, rather than technical arrogance. The latter will only foster hostility and suspicion which will guarantee that cooperation will be lacking and the fruits of the effort ignored.

Notes

1. Prepared under contract to D.C. Public Schools and under the supervision of the Division of Research and Evaluation by NTS Research Corporation (formerly IBEX, Inc.) and Roy Littlejohn Associates, Inc. (a joint venture).
2. A. Jackson Stenner, *An Overview of Information Based Evaluation: A Design Procedure*, Information Based Evaluation Series Book 1 (Durham, N.C.: NTS Research Corporation).
3. *1975-76 Final Evaluation Report*, p. 31.
4. Ibid., p. 32.

14. *Proceedings of the Practitioners' Conference*

Charles B. Stalford
Evaluation Team Leader, Testing, Assessment and Evaluation Division
National Institute of Education

The writers met for two days with TAE staff to present and discuss their papers (Judy Singleton and Parker Damon could not attend). A variety of perspectives as well as diverse backgrounds were evident among the participants at the conference.

The conference's interim objective was for the participants to receive comments on their drafts before preparing final papers for submission to NIE. The conference also provided an opportunity for supplemental discussion of themes raised in the papers. The participants were asked directly at the conference what kinds of research TAE could fund that would help meet their local needs in testing and evaluation.

The teachers presented their papers first, followed by principals and research evaluation directors. At TAE's suggestion, teachers had focused their papers on testing issues, research and evaluation directors on evaluation issues, and principals on either, as they wished. This suggestion was made in the belief that it reflected the primary interests of the respective school groups at the conference. The conference proceedings subsequently suggested, however, that the local line between testing and evaluation is often very narrow and sometimes nonexistent and, further, that these concerns are not associated predominantly with any one of the three professional groups.

For purposes of this summary, the testing issues discussed will be described first, followed by evaluation issues. Several general comments are appropriate first, however, on the interrelatedness of the topics and the concerns expressed about them.

Perhaps the most significant among these general comments is that in both testing and evaluation, participants stressed a need for practical procedural and nontechnical means to help them improve testing and evaluation activities in schools. Thus, for example, the desirability of more interaction among different local parties to testing and evaluation activities was frequently cited: teachers with principals and other administrators, practitioners of all kinds with researchers, and school officials of all kinds with the media. A more harmonious relationship among affected parties, greater mutual understanding of one another's needs, and less distortion of both the significance and limitations of testing and evaluation activities were potential benefits seen to accrue from increased interaction.

It should be noted, however, that even among the participants, satisfaction with actual experiences in such interaction varied widely. Thus, for example, while Ronald Banks reported positive results from local "participatory committees," Vic Doherty from teacher and administrator participation in formulation of district goals, and June Bland from interactive planning with teachers, Bill Moore sees the results of test writing in the Florida statewide program by a committee of teachers and administrators as disappointing. Doherty expressed a need for better training of teachers to enable them to implement local instructional objectives they have helped write.

Even Banks continues to see teachers as frequently apathetic or hostile to testing programs, despite the existence of participatory committees in Buffalo. Perhaps a useful area of research would be on conditions and processes in which participatory and interactive devices such as those recommended can best function. NIE has funded some promising research in this area.[1]

In a related vein, relatively little need was expressed for research leading to new methodological breakthroughs in testing or evaluation procedures—discussion of Rasch scaling and "edumetric testing" being exceptions; rather, much attention was directed to reducing misuse of available techniques. Concerns were expressed, for example, by Myrna Cooper, Maurice Leiter, and Ed Cypress about misuse of test results for sorting and selection. Minimum competency tests could be seen as a new methodological tool, but much discussion at the conference centered on means to cope with misuse of the technique.

In general, then, this group of practitioners was not seeking new "black box" technology of the kind so frequently publicized, if not always generally available, in medicine. A predominant orientation towards human concerns in testing and evaluation, including relationships among the adults involved as well as between the adults and students, was evident at the conference. Whether this is good or bad is a matter for the reader to judge. For the TAE staff, this orientation suggested a practical cast to new research that might be funded on testing and evaluation—not necessarily to abandon highly technical research on new methodologies, but not relying exclusively on such research to improve local practice, either.

While one might argue with some logic that "practical" suggestions from a group of practitioners for new research are not surprising, it is worth noting that the practitioners in this case were articulate and aware observers of the current education scene.

As such, they were not necessarily drawn to practically oriented research because they had no experience with anything else. This is particularly the case with the research and evaluation directors, whose greater methodological training did not, for the most part, impel them to suggest any more technical research than did teachers and principals.

Testing and evaluation can clearly be emotional as well as intellectual issues. Even in the highly professional environment of this conference, sparks sometimes flew in discussions of specific practices. Such emotions, however, only serve to underscore the importance of the issues discussed, not only for the participants, but for all in schools generally. The conference format also underscored the value of having the different groups work together, rather than separately, to deal with the issues.

Testing Issues

Testing was seen at the conference in both a political and an educational perspective. Politically, testing was seen as a manifestation of public pressures for "accountability" and more recently, minimum standards of competency.

Educationally, tests were seen by some as a potentially useful, if much abused, tool for assisting students and improving instruction. Standardized as well as minimum competency testing received a good deal of critical attention. There was one call (by a principal) for a moratorium on standardized testing but that was not shared by most of the group.

Considerable faith was expressed in the potential of criterion-referenced tests to provide instruments appropriate for the local content of what is taught and, perhaps as significantly, to assist teacher understanding of the processes as well as the product of a student's testing experience.

Considerable emphasis was given to testing that would be better linked to local instructional goals. However, even if a better link between testing and instruction could be framed theoretically, accomplishing it in practice did not seem to be a foregone conclusion. In particular, the fact that many teachers are not trained in test construction and use was cited as a problem. Teachers not versed in testing theory would be hampered in understanding and following up on such advances in testing and instruction.

At a broader level, increased interaction between teachers and other local professionals with colleges of education was seen as necessary if academic perspectives and capacities in testing were ever to be more useful in schools. Finally, getting better congruence between externally generated goals for education, such as accountability and minimum competency, with actual school instructional objectives was seen as a way to better cope with mandated testing programs. As indicated previously, a specific suggestion was to achieve a better match between the content of instructional programs and testing programs. A fear was expressed, however, that in the search for such a match, the testing "tail" might wag the instructional "dog," as some have perceived in the current "back to basics" movement.

It was suggested that NIE sponsor the following activities and research to improve testing:

- Disseminate information to local practitioners on innovative and exemplary testing practices.
- Work with the media to develop better attitudes and knowledge about testing and more responsible (e.g., less sensationalistic) reporting of results.
- Analyze the financial and human burden of testing and investigate ways to lessen it; for example, through more sampling procedures.
- Provide materials to help teachers explain test scores to parents.
- Facilitate teacher involvement in development of research activities on testing and in conduct of the research; teacher organizations might then publish abstracts of results from research on testing.
- Generate information for local practitioners on what constitutes use and misuse of tests, including appropriate ways to report results.
- Increase the link between instructional and testing research, particularly in measurement of cognitive and affective processes, and for bilingual education, and special education.
- Fund research on harmful effects of testing on minority students.

Evaluation Issues

Much of the discussion of local evaluation dealt with organizational matters and the "way it is" as opposed to the "way it's supposed to be." Thus, Blas Garza contrasted the normative organizational expectations for programmatic decisionmaking based upon sound evaluative data with the "reality" of the fragmented and hectic environment of local decisionmaking.

Organizational and personal relationships between research and evaluation staff and individual building staff as well as other "central office" personnel were seen as problematic in affecting the use of program evaluations locally. Thus, Claradine Johnson decried the lack of central office support in her effort as principal to reform a school's climate, and Michael Kean discussed the intricacies of large city school bureaucracies and how they may hinder uses of evaluation. Freda Holley highlighted the intuitively obvious if often unarticulated thought that personal relationships between researchers and school staff have much to do with acceptance of local testing and evaluation procedures.

The participants' personal observations about the importance of organizational factors in evaluation are being borne out by recent evidence from CSE's study of evaluation and decisionmaking in local districts. Where the local evaluation offices are placed—for example in the instructional division, in the administrative services division, or reporting directly to the superintendent—is seen to be associated with who the evaluation office sees as its principal clientele in the district and who uses the data.[2]

Several evaluation themes similar to those in the discussion of testing arose. Thus, the desirability of linking program evaluations to instructional objectives was cited, together with a need to find better ways of presenting results and means to work toward this end with the media. In addition, training needs were cited, in this case for evaluators to be better researchers. Many evaluators, according to Holley, come from administrative or counseling areas and lack requisite formal training in research methodology.

A corollary suggestion by Holley was for greater professional and public recognition of outstanding achievements by staff in local research and evaluation offices. Given the normal tendency of the public to accentuate the negative in this and other educational areas, Dr. Holley's suggestion may have considerable merit.

The discussion of evaluation raised two other topics: lack of money and utilization. Kean and Banks cited inadequate funding of research and evaluation offices as a problem. The claim has strong face validity; CSE's study of local evaluation offices, however, found most respondents satisfied with their level of funding.[3] This rather surprising finding suggests further investigation to see whether evaluators are actually satisfied with the funds available to them or have given up hope of obtaining more and thus secretly agree with Kean's and Banks' less sanguine feelings.

However, there is circumstantial evidence to support a claim by those concerned with all evaluation, not just with testing, that money is a problem. CSE's study and others[4] have found that most evaluation activities are *testing* programs. Title I evaluations, for example, are almost totally based on test results; available funds are therefore focused on mandated testing for Federal as well as State and local purposes. There are not frequently money or resources to perform more extensive program evaluation procedures of the kind ideally desired. Within-classroom observation by evaluators of innovative programs is rare, for example, as is support for the interactive planning procedures seen as useful by many conferees for both evaluation and testing.

This situation is related to the conferees' other unique concern regarding evaluation: low utilization. In the case of program evaluations, the villain—if that term is appropriate—perceived by local officials is the Federal Government and to a lesser extent the State governments. This is presumably because of extensive Federal funding of compensatory programs to aid disadvantaged, bilingual, special, and other distinct categories of students, each of which carries statutory evaluation requirements. The results, as highlighted by Garza, can be distressing at the local level. Federal

evaluation requirements may overlap and in some cases produce data on evaluations of different programs based on the same students!

A concomitant evil was expressed forcefully by another participant: "If it's a federally sponsored evaluation, it will be of no use locally." This claim can possibly be seen as a general rule: the further away from schools the requirement for evaluative information originates, the less useful such information will be locally.

However, this rule is not really a satisfactory explanation of the issues inherent in local utilization of Federal evaluation data and means to improve the situation. In particular, the U.S. Office of Education (USOE), spurred on by the 1978 Education Amendments, has been taking active steps to make federally required evaluations more useful locally. USOE and NIE (now both incorporated in the new Department of Education) are collaborating on a joint research project to address this and related problems in local testing and evaluation, which will be discussed in the final section of this summary.

In addition, the Holley and Bland papers provided evidence that evaluation *can* make a difference. NIE is currently using a variant of the interactionist approach to greater evaluation utilization in its "stakeholder" evaluations of the federally assisted Cities in Schools and Push for Excellence programs. "Stakeholders" are all those who have a stake in the outcome of an evaluation—program managers, parents, community leaders and policymakers, as well as funding officials. The stakeholder's strategy involves all these groups in discussions about the purposes and procedures for an evaluation at the outset and continues to engage them in communication through the life of the evaluation. It is anticipated that an evaluation so structured will be more useful to its audience than one that is not. It should be reiterated, however, that neither the papers nor the conference discussions of evaluation focused exclusively on Federal evaluation requirements. McKinley Nash, Claradine Johnson, Blas Garza, and Freda Holley discussed primarily local- and State-oriented evaluation efforts. CSE has found, in fact, that most of the budgets of local evaluation offices come from State and local sources. There is a danger, therefore, in overemphasizing the magnitude of Federal requirements. Their importance locally may lie, unfortunately, in their mandated nature.

Participants suggested NIE sponsor the following kinds of activities and research to improve program evaluations:

- Develop ways to improve presentation of evaluation results and their reporting in the media.
- Develop ways to minimize the combined burden of Federal and State evaluation requirements and eliminate overlap among them.
- Disseminate information about exemplary local evaluation practices.
- Develop ways to increase the interaction among evaluators, building staff, other administrators, and parents in order to facilitate evaluation use.
- Develop ways to better link program evaluation requirements with instructional objectives.

Next Steps

Prior to publication of these papers, TAE has already used numerous insights from the activity for its long-range planning for research on local testing and evaluation.

Procedurally, the participants were invited to review TAE's draft long-range plans in April 1979; several did so with helpful comments. Substantively, NIE and USOE initiated a joint research project on local testing and evaluation in 1979, as indicated previously. This 2-year effort being per-

formed by the Huron Institute will identify exemplary local testing and evaluation situations for analysis and dissemination to practitioners. In addition, it will identify local needs for technical assistance in testing and evaluation and explore better ways to provide such assistance.

The Methodology section of the NIE Teaching and Learning Grants Announcement for research on testing and evaluation has been significantly oriented towards the needs of local and State education agencies. Opportunities for research on such topics as how schools can better interpret individual and group test scores and how evaluations can be made more useful are provided through the Announcement. Particularly in the Evaluation area, proposals are sought from local education agencies themselves to enhance the relevance of the research to other districts. Subject to availability of funds, this grants announcement should provide a continuing opportunity at a modest level for schools to gain funds for research to improve their testing and evaluation programs.[5]

In the testing area, TAE has several other discrete projects underway that can benefit schools. A hearing based on judicial procedures to clarify the issues involved in minimum competency testing is being planned for 1981, which will be made available through videotape and written materials for national dissemination. These also include workshops for teachers on standardized testing sponsored by the American Federation of Teachers and a project to develop materials for better integrating assessment with classroom instruction. Responsive to the need for better communication to parents and others about testing, NIE is publishing a new booklet entitled, *Your Child and Testing*.[6]

In the evaluation area, research at CSE continues to be funded that will lead to better understanding of factors influencing local evaluation use. In addition, research to assist evaluation in schools is being funded at the Northwest Regional Educational Laboratory in Portland, Oregon, and the Learning Research and Development Center at the University of Pittsburgh. The first round of grants on evaluation funded through the Teaching and Learning Grants Announcement began in 1980 and should lead to further insights on needed research as well as improved practice.

TAE's plans for both fiscal year 1981 and 1982 are significantly oriented to improving testing and evaluation at the local level through its research. The activity of practitioners described herein has been a major resource in TAE's planning to improve school programs and has convinced us of the benefits to be gained from collaborative planning of Federal research programs with the constituencies to be served.

Notes

1. See William Tikunoff, Beatrice Wand, and Gary Griffin, "Interactive Research and Development on Teaching Study," Final Report (San Francisco: Far West Laboratory for Educational Research and Development).
2. See Catherine Lyon *et al., Evaluation and Decisionmaking in School Districts* (Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, December 1978).
3. Ibid., p. 60.
4. See Jane David, *Local Uses of Title I Evaluations* (Menlo Park: SRI International, July 1978) (Research Report EPRC 21).
5. *Teaching and Learning Grants Announcement*, August 1980. This announcement, available from NIE, describes opportunities for funding in fiscal years 1981-84. Address requests for copies to Teaching and Learning Grants Announcement, National Institute of Education, 1200 19th St. N.W., Washington, DC 20008.
6. Available from the U.S. Consumer Information Service, Pueblo, CO 81009.

Participants

Ronald Banks	Director of Evaluation, Buffalo, N.Y., Schools
June Bland	Division of Research and Evaluation, Public Schools of the District of Columbia
Myrna Cooper	Director, New York City Teachers Consortium, United Federation of Teachers
Edward Cypress	Teacher-in-charge, Second Grade Minischool, New York City Public Schools
Parker Damon	Principal, McCarthy-Towne Elementary School, Acton, Massachusetts
Victor Doherty	Assistant Superintendent of Evaluation, Portland, Oregon, Schools
Blas Garza	Principal, Franklin Elementary School, Santa Barbara, California
Freda Holley	Director, Research and Evaluation, Austin Independent School District
Claradine Johnson	Principal, Wichita High School East, Wichita, Kansas
Michael Kean	Executive Director, Office of Research and Evaluation, School District of Philadelphia
Maurice Leiter	United Federation of Teachers, New York, N.Y.
Luis Mercado	Principal, P.S. 75, New York City Public Schools
William Moore	Elementary School Teacher, Sanford, Florida, Schools
McKinley Nash	Principal, Evanston Township High School, Evanston, Illinois
Judith Singleton	Special Education Teacher, Fairfax County, Virginia, Schools
Catherine Lyon	Senior Research Associate, Center for the Study of Evaluation, UCLA
Mary Ann Millsap	Senior Research Associate, NIE
Corinne Scott	Senior Research Associate, NIE
Judy Shoemaker	Testing Team Leader, NIE
Charles Stalford	Evaluation Team Leader (Chair), NIE

*U.S. GOVERNMENT PRINTING OFFICE: 1980-0-722-206/367